The Structure of Complex Networks

Jørgensen, Sune Lehmann

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

The Structure of Complex Networks

Sune Lehmann Jørgensen

Kgs. Lyngby
February 2007
IMM-PHD-2007-176

Contents

List of Figures

Abstract

THIS dissertation regards the structure of large complex networks. The dissertation is divided into three main parts. Part I contains chapters 1–3. Part II contains chapters 4–6. Part III consists of the concluding chapter 7 and the bibliography.

Part I serves as an introduction. Chapters 1 and 2 directed toward enabling a general reader to understand the concepts and nomenclature used in the research, presented in Part II of the dissertation. These chapters also motivate and explain the unifying idea behind the research contained in this dissertation. Chapter 3 describes the origin and general structure of the data set used in the subsequent chapters.

Part II contains five papers that chronicle my research as a Ph.D.-student.

- **Chapter 4** consists of two papers: *Life, Death, and Preferential Attachment* [54] and *Live and Dead Nodes* [53] where a mathematical model of the network of scientific papers is motivated empirically and solved analytically. The model is an augmentation of the growing networks model, first introduced by Barabási and Albert [10]. In [53, 54], the idea that network nodes can 'die', is introduced, and the associated consequences for the growth model are explored. Further, it is demonstrated that the mechanism for 'node-death', alone, can create networks with power-law degree distributions.

- **Chapter 5** concerns the longitudinal correlations, in the citation network, that is induced by the authors of the scientific papers. This chapter also contains two papers: *Measures for Measures* [56] and *A Quantitative Analysis of Measures of Quality* [57]. Here, Bayesian statistics are employed to analyze several different measures of quality. Using scaling arguments, it is demonstrated how many papers are needed to draw conclusions, regarding long-term scientific performance with usefully small statistical uncertainties. Further, the approach described here permits the value-free (i.e., statistical) comparison of scientists working in distinct areas of science.

- **Chapter 6** discusses the detection of communities in large networks, using deterministic mean field methods. Further, the paper, *Deterministic*

Community Detection [52] presents an analytical analysis of a simple class of random networks, with adjustable community structure.

Part III consists of a final, concluding chapter that recapitulates the main ideas, presented in this dissertation, and points towards new avenues for further research. Additionally, part III also contains the bibliography.

Resumé på dansk

MIN forskning drejer sig om at forstå strukturen af store komplekse netværk. Denne afhandling indledes med tre kapitler der introducerer feltet og datamaterialet. Mit eget arbejde har i særlig grad drejet sig om tre områder. Artiklerne *Life, death, and preferential attachment* [54] og *Live and dead nodes* [53] omhandler analysen af en simpel netværksmodel, der er karakteriseret ved, at knudepunkterne i denne model kan blive inaktive (dvs. 'dø'). Artiklerne *Measures for measures* [56] og *A quantitative analysis of measures of quality in science* [57] beskriver en Bayesiansk analyse af netværket af forfattere i SPIRES-databasen for højenergi partikelfysik. Denne analyse benyttes blandt andet til at vurdere pålideligheden af forskellige mål for videnskabelig kvalitet. Sluttelig, drejer artiklen *Deterministic community detection* [52] sig om at forstå den modulære stuktur som nogle netværk besidder. Ved hjælp af middelfeltsmetoder, fx kendt fra spin-glas modeller i fysik, bestemmer vi modulerne i en simpel klasse af Erdös-Rényi netværk med indbygget modulær struktur. Vi har opnået en analytisk forståelse af disse netværk og kan derfor variere deres modularitet.

Acknowledgements

I acknowledge that I am eternally indebted to a number of people for help and inspiration. Andy Jackson and Benny Lautrup have graciously dispensed experience and expertise regrading both science and life in general—not to mention that they have dramatically expanded my repertoire of dirty jokes. As my advisor, Lars Kai Hansen has been able to strike a perfect balance, showing trust in my choices and generously giving me the freedom to pursue them, while always being there to support me when that was needed.

I acknowledge that I am a lucky person. This is due to my family: I have inherited a positive outlook on life and an inability to feel depressed for more than fifteen minutes at a time—life always supplies a reason to crack a smile. My family has always met me with complete trust and unconditional support; I have always known that this support and trust will follow me in any endeavor I choose to pursue. I am thankful for all of these gifts every moment of my life.

Aino is where my world begins and ultimately ends. She is beautiful, independent, and wise. I love her more than I could ever begin to explain! I hope and strive to maintain her love and affection, always.

Part I

A History of Structure

CHAPTER 1

Models and Measures

COMPLEX networks are everywhere. Every living person is endlessly entangled in myriads of webs: Sociological networks of friends, coworkers or sexual partners; information networks, such as the world wide web, phone lines, and the internet[1]. We travel on transportation networks, such as roads, trains, or airplanes; and participate in biological networks ranging from macroscopic food-webs to microscopic regulatory networks that exist inside each of our cells.

The existence of these networks has been known to anyone with an active imagination since the beginning of time, and the odd and surprising links between people and places have been explored in art and literature for centuries. It is, however, only with the advent of the personal computers, the world wide web, and easy access to advanced data storage systems, such as relational data bases, that data sets regarding any of these networks have become available.

Large data bases of observations of these diverse real world networks form the foundation of the science of complex networks. Since the theoretical under-

[1] By the word 'internet', I mean the physical system that enables the existence of the world wide web.

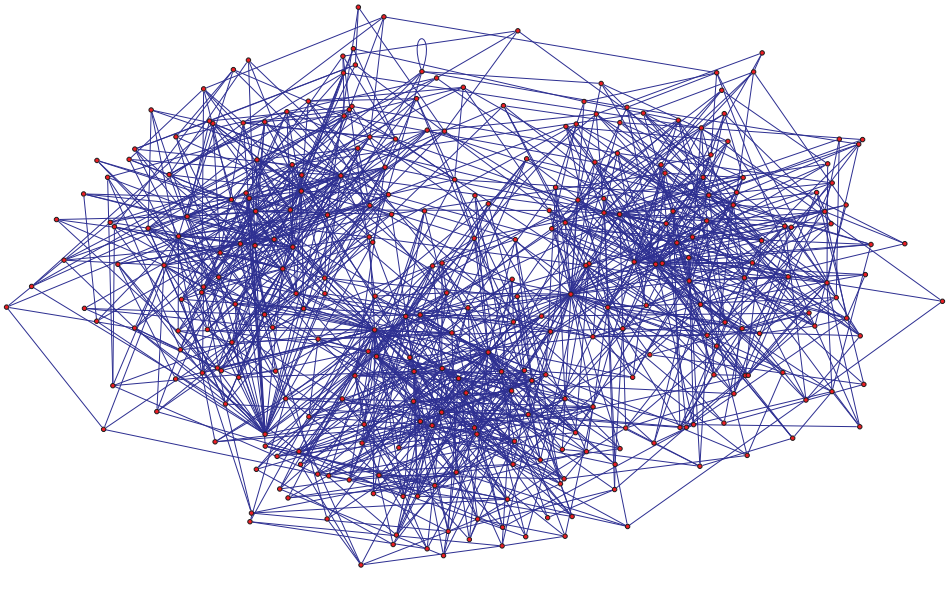


Figure 1.1: A network of 300 nodes, which is approximately the largest amount of nodes it is possible to plot on a piece of paper without the links being an indistinguishable solid blue surface on the page. This network has 3 communities and a scale free degree distribution. The positions of the nodes are determined using an algorithm developed by Kamada and Kawai [40] and the network is plotted using the program *Pajek* [92].

standing of large complex networks is still in its infancy, much of the research in this field has focused on the development of statistical tools necessary to understand and categorize the richness and diversity present in the complex networks that surround each one of us.

As an illustration of the problems we are facing, figure 1.1 displays a small network of 300 nodes. From this figure we can gain an idea of the network structure. It is clear to see that the network has 3 communities, it may also be possible to observe a large variation in the number of links connected to each node: Some nodes have only a few links to other nodes, while a few nodes have many more. The image also supplies some information about the typical path-length of the network—due to the existence of the hubs, all nodes can be reached from anywhere in a small number of jumps.

Most analyses of real world networks in the previous century, proceeded along

these lines. Given a picture of a network, we can ask questions about its structure and find the answers by visual inspection. In fact, through the course of evolution, our brain has become exceptionally skilled at analyzing information in the visual field (for instance, see if you can spot the loop in figure 1.1). With a network of a million or a billion nodes, however, the 'eyeballing' approach described above is completely useless. Considering how unruly a network of 300 nodes looks on the page, imagine the difficulty of analyzing networks that are just 10 times larger; for even larger networks, the initial task of determining an (x,y)-position for each node alone can bring super-computers to their knees [40].

This is precisely the reason we use mathematics to study network structure! The goal is to find statistical methods that can help to inform us what a particular network 'looks like' even though we cannot actually observe it. The attempt to overcome this challenge is the motivation of the work described in the remainder of this dissertation. In this and the following chapter, I will outline the origins of the science of complex networks with a focus on how successive discoveries of surprising structural properties has shaped our understanding of these massive data sets. In order to make things as simple as possible, I will focus (almost[2]) exclusively on networks where the links are *undirected* and *unweighted*; many results can, however, easily be generalized to the directed and weighted cases. For alternatives to the general review of the development of network science presented below, the reader is referred to [5,9,13,14,23,65,71,93].

1.1 Random Networks

In terms of both chronology and structural complexity, *random networks* are the starting point. Random networks are designed to possess no structure. A random network is entirely defined by its number of nodes n, and the probability p that a link exists between each pair of nodes. Random networks have been an active research area in mathematics for many years and many of their properties are known analytically. The field of modern graph[3] theory was started by Erdös

[2]Whenever networks are directed or weighted, this will be stated explicitly in the text.

[3]The word 'graph' is often used instead of 'network'—especially in the mathematics literature.

and Renyi [25].

It follows immediately from the definition of a random graph, that when n is large, the random network has $m = pn(n-1)/2$ links. The number of links to and from node i is denoted k_i and called the *degree* of node i. The average degree of a node in a random network, is $\langle k \rangle = np$. Having defined the degree of a node, we can also consider the distribution of degrees. The *degree distribution* is simply a function describing the number of nodes $N(k)$ with a certain degree k. The normalized degree distribution describes the probability $P(k)$, that a node in the network has degree k. In the following, we will adhere to the vernacular of network science and designate the words 'degree distribution' to describe the normalized degree distribution. The degree distribution of random networks follows the Poisson distribution, thus

$$P(k) = \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!}. \tag{1.1}$$

1.2 Watts and Strogatz

The first clue, that the random network model is too simple to be used as a description for real world networks, came from the *clustering coefficient*. Watts and Strogatz were attempting to gain a quantitative understanding of an earlier result, by the socio-psychologist Stanley Milgram. In a simple experiment, Milgram [61] documented that in the social network where links are constituted by human relationships (friends and acquaintances), the average distance between two nodes is very short. On average six steps are enough to connect any two people in the western hemisphere. Networks that possess this quality are called *small world* networks.

A random network is a small world network. An average node in a random network has $\langle k \rangle$ neighbors, $\langle k \rangle^2$ second neighbors, $\langle k \rangle^3$ third neighbors, etc. If a typical person has 100 friends and acquaintances, six steps will result in an extended network of 10^{12} persons. Since only 6 568 751 041 people inhabit the globe[4], this is more than enough to account for Milgram's result.

[4] According to the U.S. Bureau of the Census, this was the world population aon January 9th 2007 20.27 GMT (time of writing).

The problem with random networks as a model for social networks, is the fact that real social networks display a property called *clustering*. Loosely speaking, clustering describes the propensity of a person's friends to also be friends with each other. In a more technical formulation, social networks tend to have a much higher percentage of triangles than random networks, since a triangle is the network manifestation of two neighboring nodes being connected. Watts and Strogatz [94] define the clustering coefficient c_i of node i, as the actual number of links between node i's neighbors divided by the maximum number of links that could exist between them

$$c_i = \frac{2t_i}{k_i(k_i-1)} \quad \text{and} \quad C = \frac{1}{n}\sum_{i=1}^{n} c_i\,, \tag{1.2}$$

where t_i is the actual number of links, (or, equivalently, the number of $t$riangles) that i participates in, and $k_i(k_i-1)/2$ is the maximum possible number of links between node i and its neighbors. The network clustering coefficient C, is defined as the average of all node clustering coefficients.

First of all, note that the presence of clustering implies that the assumption of a node having $\langle k\rangle^2$ second neighbors, breaks down. This is due to the fact that because of the triangles, many of the second neighbors have already been counted as his first neighbors. Second of all, random networks display almost no clustering. No part of the definition of random networks encourages the formation of triangles. It turns out that for random networks with $p < 1$, $C_{rand} \approx O(n^{-1})$ [94]. This clustering coefficient is much lower than that of most real world networks.

The solution that Watts and Strogatz propose is based on another very simple network that displays abundant clustering: The regular lattice. A d-dimensional regular lattice where each node is connected to its k nearest neighbors, has the clustering coefficient

$$C = \frac{3(k-2d)}{4(k-d)}, \tag{1.3}$$

which tends to $3/4$ for $k \gg 2d$. Low dimensional lattices do not display the small world property. Unless k or d is large compared to the total network size n, many jumps are needed to get from one side of the network to the other. But, Watts and Strogatz demonstrated that all we need is 'a little randomness'. If

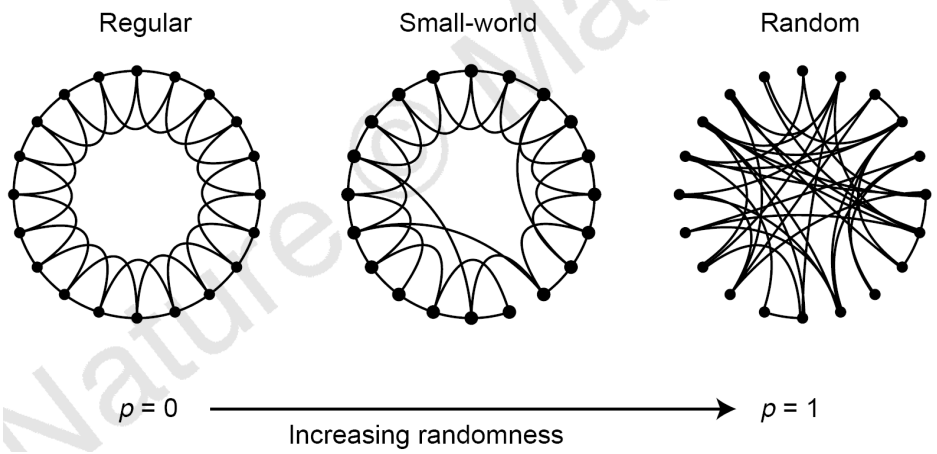


Figure 1.2: The Watts-Strogatz idea of network structure illustrated on a one dimensional lattice with $k = 4$. On the left is the regular lattice; this network is clustered but does not exhibit the small world property. In the middle, we see the Watts-Strogatz model. When each link is rewired with probability $p' = 1$, we regain the random network, which is displayed on the right. The small world property emerges for even very small p'. Image from [94].

we, with some small probability p', rewire each link randomly, the small world property emerges [94].

The Watts-Strogatz model manages to capture some structural elements of social networks that correspond very well to our intuitions. We all have a group of close friends that we interact with on a day-to-day basis. Also, we all have friends where the link is further ranging in time, place, or social stratus—that one childhood friend that we are still in contact with, or someone we have met while traveling, etc. Those links are what makes our world a small one.

There is only one problem with the Watts-Strogatz model. It makes one key assumption about network structure that is completely wrong! It assumes that the degree distribution of real world networks is roughly Poissonian[5]. In his book about complex networks [93], Watts regrets this mistake:

> "*We didn't check!* We were so convinced that non-normal degree distributions weren't relevant that we never thought to look at which networks actually *had* normal degree distributions and which did not. We had the data sitting there staring at us for almost two years,

[5]For large networks, the Poissonian degree distribution from equation (1.1) converges towards the normal distribution.

> and it would have taken all of half an hour to check it, but we never did."

It is easy to understand Watts' frustration. The mistake of assuming a Gaussian degree distribution, illustrates precisely why the study of structure in complex networks is important. Without the proper statistical tools to analyze these massive networks, our background as statistical physicists can lead us to believe that the assumptions of homogeneity and simplicity that have been successful in the study of gasses and solids, might automatically be valid for networks. As we shall see in the following, this is *not* the case.

1.3 Hubs and Power-laws

Barabási and Albert discovered that the degree distribution of many real world networks is far from normal; they discovered that *hubs* are an essential part of networks. In an important paper from 1999 [10], they document that the degree distribution of many real world networks follows a power-law, that is

$$P(k) \sim k^{-\gamma}, \tag{1.4}$$

where $\gamma > 1$; typically $2 < \gamma < 3$. In figure 1.3, the cumulative degree distributions[6] for several real world networks are plotted. Panels (c), (d), and (f) of this figure display power-law behavior over many orders of magnitude and panels (a) and (b) display asymptotic power laws (no power-law for small k). To emphasize that not all networks have power-law degree distributions, panel (e) shows one network where the data decays according to an exponential curve.

Probability distributions with asymptotic power-law behavior are very different from the familiar normal-type distributions (such as the Poisson- and Gaussian

[6]When plotting power-laws, it is often a good idea to use the cumulative distribution. Since the cumulative distribution stems from the integral of the original distribution, the cumulative distribution of a power-law degree distribution also displays a straight line on a log-log plot. The cumulative distribution does not need to be binned for plotting, and because binning can often be a problem when plotting the tail of power-laws directly (because of the sparsity of data in this region), the cumulative plot is often a good choice for plots. See [67] for a detailed discussion of these matters.

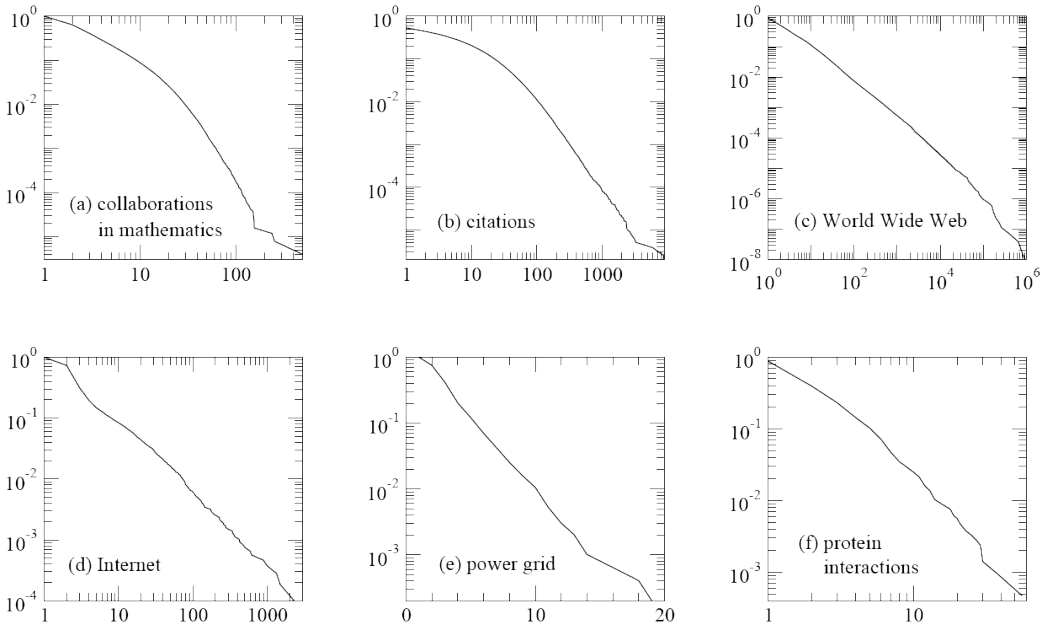


Figure 1.3: Various (asymptotic) power-laws and one exponential. This figure displays the cumulative degree distributions for six different networks. The x-axis for each panel is degree k and the y-axis is fraction of vertices that have degree greater than or equal to k. The networks shown are: (a) the collaboration network of mathematicians; (b) citations between 1981 and 1997 to all papers cataloged by the Institute for Scientific Information; (c) a 300 million vertex subset of the World Wide Web, circa 1999; (d) the Internet at the level of autonomous systems, April 1999; (e) the power grid of the western United States; (f) the interaction network of proteins in the metabolism of the yeast S. Cerevisiae. Network (e) has an exponential degree distribution (note the log-linear scales used in this panel). Data and figure from [65].

distributions) that are centered around some typical value and decay exponentially for values both larger and smaller than the mean. Power-law distributions are often called *scale free distributions*; we will address the frequent (mis)use of this term below.

In order for a power-law to be normalizable, it must have a non-zero minimum value $k_{\min}$.[7] This is the most probable number of links. From the maximum value $P(k_{\min})$ the distribution decays towards zero for $k \to \infty$; but the rate with which it decays is much slower than the rate of normal-type distributions. This *heavy tail* of power-law probability distributions, results in the fact that extreme events are more likely; therefore, there is a large factor difference between the mean and the median number of nodes in a scale free network, because the value of the mean is highly influenced by these extreme events.

We can quantify this by calculating the moments of a normalized power-law distribution. The normalization constant is determined by

$$1 = C \int_{k_{\min}}^{\infty} P(k)\, dk = \frac{C}{1-\gamma} \left[x^{-\gamma+1} \right]_{k_{\min}}^{\infty} \tag{1.5}$$

which yields $C = (\gamma - 1)k_{\min}^{\gamma-1}$. Here, we also see explicitly why the power-law probability distribution is only defined for $\gamma > 1$, the integral diverges for $\gamma \leq 1$. The mean value of a power-law is calculated along similar lines

$$\langle k \rangle = \int_{k_{\min}}^{\infty} kP(k)\, dk = \frac{C}{2-\gamma} \left[x^{-\gamma+2} \right]_{k_{\min}}^{\infty}. \tag{1.6}$$

This integral becomes infinite for $\gamma \leq 2$. Thus, for a power-law with slope smaller than or equal to two, no finite mean exists. Any finite sample from this distribution will, of course, have a finite mean but, as we allow the sample size grow, we have a non-neglible probability of getting a larger maximum value for the set. This value will increase the mean. For larger and larger data sets the mean will increase without bound. If $\gamma > 2$, the value of $\langle k \rangle$ will settle down to a finite value as the data set becomes large. In the case of network analysis this

[7] In practice, the value $k = 0$ is often important. Some networks have a large number of nodes with zero links. When plotting the degree distribution of these networks, one must take this fact into account. A common solution to this problem is to use $k + 1$ for the plots instead of k.

is not a significant problem, since most real world networks have slopes that are greater than two. Inserting the limits and the constant C, we find that

$$\langle k \rangle = \frac{\gamma - 1}{\gamma - 2} k_{\min}. \tag{1.7}$$

We need to be even more careful with the second moment; this is due to the fact that the integral

$$\langle k^2 \rangle = \int_{k_{\min}}^{\infty} k^2 P(k)\, dk = \frac{C}{3-\gamma} \left[x^{-\gamma+3} \right]_{k_{\min}}^{\infty}, \tag{1.8}$$

diverges for $\gamma \leq 3$. In other words, the variance of the mean is undefined for networks with slope $\gamma \leq 3$. Since γ is typically in the range from two to three, this poses a serious problem. If the second moment is well defined, we can again insert the limits and the constant to find

$$\langle k^2 \rangle = \frac{\gamma - 1}{\gamma - 3} k_{\min}^2, \tag{1.9}$$

which complements equation (1.7).

In summary, we should exhibit great care when using the mean degree to gain knowledge about a network with a power-law degree distribution. In the case of networks with power law degree distributions, the mean is highly influenced by the tail of the distribution. This means that only a small fraction of the network nodes will have degrees higher than the mean degree. Further, if the slope of the power-law is smaller than or equal to $\gamma = 3$, then the variance of the mean is undefined. An undefined variance leads to the consequence that finite samples of nodes, from such a distribution, will have a diverging mean and will contain, essentially, no information about the original distribution. This fact will turn out to be important in chapters 4 and 5, when we study the network of scientific citations and references. An author's citation record is precisely a small sample of nodes from a network with a power-law degree distribution. Using the mean to characterize such a sample could therefore be highly problematic.

A more balanced alternative to the mean is the *median*. The median is well-defined for any power-law distribution with $\gamma > 1$. For a normalized distribution, it is defined as the point $k_{1/2}$ where the distribution is divided in two

$$\int_{k_{\min}}^{k_{1/2}} P(k)\, dk = \frac{1}{2}. \tag{1.10}$$

We can solve this equation for $k_{1/2}$ to find

$$k_{1/2} = 2^{1/(\gamma-1)} k_{\min}. \tag{1.11}$$

The strength of the median (as opposed to the mean) is that it has this clear and intuitive interpretation for a power-law probability distribution. This is the case for any percentile measure. For example, we can find the number of links $k_{9/10}$ that a node needs to possess in order to be in the top ten percent of the most connected nodes. These 'percentile measures' have the further advantage that their variances are always well defined [56].

Power-law distributions are often referred to as *scale-free distributions*. This is because a power-law distribution looks the same on any scale. We can formalize this statement by defining a scale-free distribution as a distribution $q(x)$ that satisfies the criterion

$$q(ax) = f(a)q(x) \tag{1.12}$$

for any a [67]. In plain words equation (1.12) simply states that if we increase the *scale* by which we measure x by a factor of a, the overall shape of the distribution is unchanged, except for a multiplicative constant. The power-law distribution is the only distribution that fulfills this criterion. Let us see why this is the case.

Setting $x = 1$, we find that $f(a) = q(a)/q(1)$. We now write equation (1.12) as

$$q(ax) = \frac{q(a)}{q(1)} q(x). \tag{1.13}$$

Since this must be true for any choice of a, we can differentiate both sides of the equation with respect to a and find

$$x\, q'(ax) = \frac{q'(a)}{q(1)} q(x), \tag{1.14}$$

where the prime denotes the derivative of q with respect to its argument. Inserting $a = 1$ yields

$$x \frac{\mathrm{d}q}{\mathrm{d}x} = \frac{q'(1)}{q(1)} q(x). \tag{1.15}$$

which is a simple first order differential equation. The solution is

$$\ln q(x) = \frac{q(1)}{q'(a)} \ln x + c\,. \tag{1.16}$$

We can find the constant $c = \ln q(1)$ by inserting $x = 1$. All we have left now is exponentiating both sides. This yields

$$q(x) = q(1)\, x^{\alpha}, \tag{1.17}$$

where $\alpha = q(1)/q'(1)$. In other words, in order to fulfill the scale-free criterion, our $q(x)$ must be a power-law distribution.

In this sense, the use of 'scale-free' to describe any distribution that exhibits power-law behavior in only a part of its domain is a clear misnomer. Any 'break' in the power-law introduces a typical scale. Strictly speaking, an empirical distribution should therefore only by denoted 'scale-free' if it is described by a power-law over the entire domain of x-values. This is not the case for most real world degree distributions; they typically have an asymptotically scale-free tail. In the scientific network literature, however, it is common to use the expression 'scale-free' to denote almost any broad distribution, and in the following, I will sometimes use the expression in this—less strict—sense.

It is worth mentioning that the concept of scale-free distributions plays an important role in modern statistical physics. One example is the correlation length (a typical scale) of magnets. For certain *critical* values of the governing parameters, the system undergoes a phase transition, where the correlation length diverges and become scale-free. The argument given above implies that at such a point the observable quantities in the system should adopt a power-law distribution.

Thus far, we have adhered to the tradition in network theory and analyzed the power-laws under the tacit assumption that the degree distribution is real-valued and positive. Since the degree distribution, in reality, is discrete, it is interesting to take a look at power-law distributions for discrete variables; for now, we will therefore assume that k is defined only on the integers. One way to proceed is to declare that k follows a power-law if

$$P(k) = C_1 k^{-\gamma}. \tag{1.18}$$

Since equation (1.18) diverges for $k = 0$, the smallest possible value for k is $k = 1$. Therefore, the normalization is given by

$$1 = \sum_{k=1}^{\infty} P(k) = C_1 \sum_{k=1}^{\infty} x^{-\gamma} = C_1 \zeta(\gamma), \tag{1.19}$$

where $\zeta(\gamma)$ is the Riemann ζ-function. Thus, $C_1 = 1/\zeta(\gamma)$ and

$$P(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}. \tag{1.20}$$

Most of the results shown above can be generalized to discrete variables in this way—although the results often require special functions instead of the more tractable integrals we have encountered so far.

For reasons that will become clear shortly, a better definition of a power-law for discrete variables obtained by replacing equation (1.18) by

$$P(k) = C_2 B(k, \gamma), \tag{1.21}$$

where $B(a, b)$ is the Legendre B-function[8] given by

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{1.22}$$

For large a and fixed b, we have that $B(a, b) \sim a^{-b}$. This result can be derived by applying Stirling's approximation to the Γ-functions. The distribution given by equation (1.21) is called the *Yule*-distribution, after the Scottish statistician Udny Yule who derived it in 1925 [97]. The Yule-distribution is especially satisfying because many of the sums used for normalization and the different moments can be calculated in closed form.

Since the Γ-function diverges for $k_{\min} = 0$, the smallest possible value is again $k_{\min} = 1$. In this case, the normalization is given by

$$1 = C_2 \sum_{k=1}^{\infty} B(k, \gamma) = \frac{C_2}{\gamma - 1}, \tag{1.23}$$

which yields $C_2 = \gamma - 1$ and

$$P(k) = (\gamma - 1) B(k, \gamma). \tag{1.24}$$

If the degree distribution has a $k_{\min}$ higher than 1, we find that

$$1 = C_3 \sum_{k=k_{\min}}^{\infty} B(k, \gamma) = C_3 B(k_{\min}, \gamma - 1), \tag{1.25}$$

[8] The 'B' is a capital β, so the pronunciation is 'beta-function'.

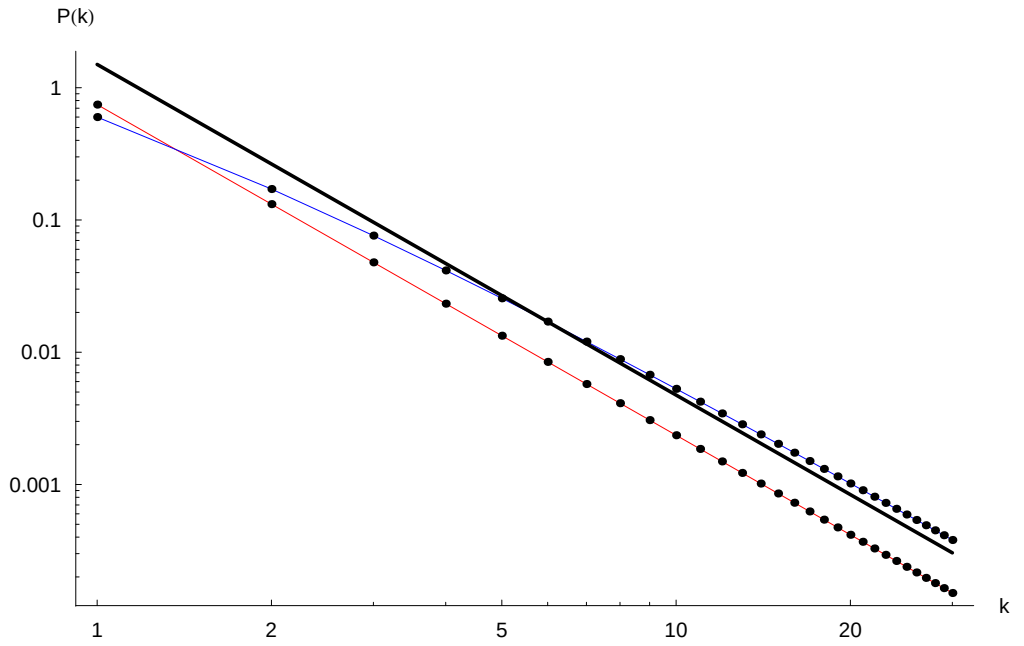


Figure 1.4: Discrete and continuous distributions. The figure displays $P(k)$ for $k \leq 30$ for three normalized power-law probability distributions with $k_{\min} = 1$ and slope $\gamma = 2.5$ on log-log axes. The solid black line is a continuous function of the type seen in equation (1.4). Note that in this case, $P(1) > 1$. The black dots connected by a thin red line for visual guidance are the normalized discrete expression from equation (1.20); this line is parallel to the continuous distribution on the log-log plot, but normalized on the integers. Finally, the black dots connected by a blue line follow the Yule distribution from equation (1.24). The Yule distribution does not follow a power-law for the lowest values of k, but has many desirable qualities—for example the first moment of the Yule distribution is identical to the first moment of the continuous power-law probability distribution.

and, the fully normalized discrete power-law function becomes

$$P(k) = \frac{B(k,\gamma)}{B(k_{\min},\gamma-1)}. \tag{1.26}$$

In the simple case, where $k_{\min} = 1$, the first moment is given by

$$\langle k \rangle = \frac{\gamma-1}{\gamma-2}, \tag{1.27}$$

This expression is only valid for $\gamma > 2$. Similarly, the second moment is only

defined for $\gamma > 3$ and given by

$$\langle k^2 \rangle = \frac{(\gamma-1)^2}{(\gamma-2)(\gamma-3)}. \tag{1.28}$$

These two expressions complement equations (1.7) and (1.9) nicely. We will encounter the Yule-distribution frequently in the following.

With the quantitative considerations behind us, we can begin to consider some of the more qualitative implications of the power-law degree distributions: What influence does the fact that most nodes in a network are not very connected and a select few are very well connected have on the network properties? As it turns out, the existence of *hubs*[9] has a profound impact on most network properties. The small-world effect that Watts and Strogatz had to design arises spontaneously in many scale-free networks [19]. This is easy to understand: Hubs act as an equivalent of long range links in the Watts-Strogatz model. Think about the case of the scale-free network of air-traffic, a hub is always close and a hub can get you close to where you are going.

The presence of hubs have many other important implications: A classical reason for studying networks is understanding the spreading of infectious diseases: it turns out that the dynamics of epidemics is vastly different if the network the disease is spreading on has a scale free degree distribution [79]. In a related vein, search strategies are radically different on networks with a power-law degree distribution; this fact affects many entreprises from web-search [16, 43] to peer-to-peer (P2P) file sharing [2, 3, 41]. Another important example is networks' vulnerability to attacks and failures; scale-free networks are much more resilient to random attacks than networks with normal-type degree distributions, but they are much more vulnerable to attacks of highly connected nodes [7].

In the next section, we will discuss how these hubs can come into existence and the our study of network structure will take a different direction, when the

[9] In a network with a heavy tailed degree distribution, some vertices have a degree that is orders of magnitude larger than the average: Due to heavy advertising especially by Barabási [9], these vertices are often called 'hubs', although this expression is a bit misleading as there is no inherent threshold above which a node can be viewed as a hub. If there were, then it would not be a scale-free distribution!

time-evolution of networks is used to explain the power-law degree distribution discussed here.

1.4 Growth Models

Barabási and Albert [10] did not just document that many real world networks have power-law degree distributions, they also suggested plausible mechanisms[10] by which power-law degree distributions can come into existence. Instead of designing an entire network, Barabási and Albert suggested two mechanisms that together result in networks with power-law degree distributions.

The first of these two mechanisms is *growth*. We imagine a network that grows one node at at time. At each update the new node links to a number of nodes that are part of the existing network. The second mechanism is *preferential attachment*. Preferential attachment states that new nodes link to existing nodes with a probability proportional to their degree. Together, these two mechanisms constitute the growth-model. In order to set up a simulation for this model, we need to make a number of choices: Should we simulate a directed or an undirected network? How many nodes in the existing network, should each new node link to? Do we want weighted or unweighted links? Is it necessary that the preferential attachment is precisely proportional to the degree k of the nodes in the network or could it be proportional to k raised to some power? Etc.

Here, we will begin by considering a simple model where each new node links to ℓ nodes in the existing network. As we have assumed the previously our network will be unweighted and undirected. Since our network is undirected, the initial probability of a node acquiring a link is proportional to $k = \ell$, because it enters the network with that number of links.

There are several ways to find an analytical solution for the degree distribution of this model, but we will apply a variation of the rate equation approach for-

[10] This mechanism has been 'discovered' at least twice before. Herbert Simon discovered a mechanism for generating power-laws as early as 1955 [89]. Simon's work was rediscovered by de Solla Price in 1976 [21], who was the first to use these ideas in the context of networks (of scientific citations). Price also found an analytical solution to the model. In this sense, the Barabási-Albert model should arguably be called the Price-model. In the following, I will call it the 'growth-model'.

mulated by Krapivsky *et al.* [49]. As above, $P(k)$ is the probability of finding a node with k links. The rate equations are

$$P(k) = \ell \left[\lambda_{k-1} P(k-1) - \lambda_k P(k) \right] + \delta(k, \ell), \tag{1.29}$$

where the λ_k are rate constants. The preferential attachment is included in the rate constants by:

$$\ell \lambda_k = ak. \tag{1.30}$$

The first term on the right hand side of equation (1.29) thus accounts for nodes with $k-1$ links gaining a new link, and the second term on the right hand side accounts for nodes with k links being bumped up to $k+1$ links. The δ-function accounts for the new node with ℓ links. We define $P(k)$ to be zero for $k < \ell$ and since all nodes must have a finite number of links, the $P(k)$ must become exactly zero for sufficiently large k. Thus all sums run from $k = \ell$ to infinity.

The $P(k)$ in equation (1.29) trivially satisfy the normalization condition

$$\sum_{k=\ell}^{\infty} P(k) = 1. \tag{1.31}$$

Equation (1.29) must also satisfy the constraint from the mean number of links. Since all nodes are loaded with ℓ links, and the network is undirected (which means that all links are counted twice), we must have

$$\sum_{k=\ell}^{\infty} kP(k) = 2\ell. \tag{1.32}$$

We must impose this constraint by an overall scaling of the λ_k. Solving the recursion, we find

$$P(k) = \frac{B(k, \eta)}{B(\ell, \eta - 1)}, \tag{1.33}$$

which is simply the Yule distribution normalized on the interval from ℓ to infinity, where we have introduced $\eta = 1 + 1/a$, cf. equation (1.26). But, the constraint from the mean in equation (1.32), forces us to set $a = 1/2$ and in turn $\eta = 3$. The fact that η is an integer allows us to write the Γ-functions as factorials and leads to a tremendous simplification. We find that

$$P(k) = \frac{(\ell+2)!}{2!(\ell-1)!} \frac{(k-1)!2!}{(k+2)!} = \frac{2\ell(\ell+1)}{k(k+1)(k+2)}, \tag{1.34}$$

since most of the products in the factorials cancel out. In spite of the simplicity of equation (1.34), this formally remains a special case of the Yule-distribution. The solution is valid for all $k \geq \ell$ and, for large k, it clearly scales like $P(k) \sim k^{-3}$.

Familiarity with the special case above, where the constraint from the mean in equation (1.32) simplified the solution greatly, clearly points towards a way to create networks with a more complex degree distributions. As we know from our study of the Yule-distribution, the unconstrained solution in equation (1.33) would allow for a fit to almost any power-law distribution. In the following, we will add a little more wiggle-room to the model by considering a directed network and adding a new parameter to obtain a more flexible model.

In a directed model, each node is introduced to the network with ℓ out-links. Thus the out-link degree distribution is simply

$$P(k^{(\text{out})}) = \begin{cases} 1, & \text{if } k^{(\text{out})} = \ell \\ 0, & \text{otherwise.} \end{cases} \tag{1.35}$$

For this reason, we want to model the in-degree distribution—in order to simplify the notation, let us simply use the notation $k^{(\text{in})} = k$ to denote the number of in-links. Since we distinguish between in- and out-links, we have a problem initiating the preferential attachment: Each new node has no in-links and therefore no 'attraction'. This can be remedied by introducing a parameter k_0 which denotes a number of 'ghost in-links' that we use to initiate the preferential attachment. The k_0 can be removed after running the model. Note that k_0 can also be used to tune when the preferential attachment 'kicks in'; for $k_0 \to \infty$ there is no preferential attachment in the model. The rate equation for the directed network is given by

$$P(k) = \ell \left[\mu_{k-1} P(k-1) - \mu_k P(k) \right] + \delta(k,0), \tag{1.36}$$

where the new rate constant is given by $\ell \mu_k = a(k + k_0)$ and the normalization conditions are

$$\sum_{k=0}^{\infty} P(k) = 1, \qquad \sum_{k=0}^{\infty} k P(k) = \ell. \tag{1.37}$$

Again, the constraint from normalization is trivially fulfilled for all choices of μ_k, and the constraint from the mean must be imposed by an overall scaling of

the rate constants. Note that the mean number of in-links in this model is equal to ℓ since the total number of out-links must match the total number of in-links.

This time, solving the recursion and imposing the constraint yields

$$P(k) = \frac{B(k + k_0, \beta)}{B(k_0, \beta - 1)}, \tag{1.38}$$

for the in-degree distribution. Again, we have introduced a new constant $\beta = 2 + k_0/\ell$. This model does indeed supply us with more freedom. We can now 'tune' the slope of the asymptotic power-law to any number greater than 2 by adjusting the values of k_0 and ℓ.

During the first years of this millennium, much of the work performed in the complex network community was centered around models such as the two presented above—gaining a deeper understanding of the analytical aspects and setting up more complex models in order to emulate real systems in greater detail. It was assumed that, if a network had a power-law degree distribution, then some kind of preferential attachment was the source. Later, however, it has become clear that many other mechanisms can cause power-law degree distributions, see [67] for an overview. Today the consensus is that some variation of the growth model gives a good description of the world wide web-network and the network of scientific citations. As we shall see in the following, even in the case of these two networks, this model runs into problems.

Chapter 4 of this dissertation concerns modeling of the network of scientific citations and references. The problems of the simple growth model in that context will taken up there (see also section 3.3). Let us therefore consider the case of the internet. As it turns out, the problems are again related to the inability of the model to capture important elements of the network structure. Since we are not 'designing' these growth models in the sense that we were designing the Watts-Strogatz networks, we must first simulate or solve analytically the network and then analyze the structure. In an analytic study of the growth model, Krapivsky and Redner [47] showed that this model has two important types of correlations. Firstly, these authors noted a correlation between the degrees of adjacent nodes.

Secondly, and more importantly in this context, there is a correlation between

age and degree of a node, with older nodes having a higher mean degree than nodes more recently added. In the simple case of Eq (1.34) with $\ell = 1$, we have that the probability distribution of the degree of a node i with age α (counted in number of nodes added after node i) is given by

$$P[k(i,\alpha)] = \sqrt{1-\frac{\alpha}{n}}\left(1-\sqrt{1-\frac{\alpha}{n}}\right)^{k}, \tag{1.39}$$

Thus, for specified age α, the distribution is exponential with a characteristic degree scale that diverges as $(1-\alpha/n)^{-1/2}$ as $\alpha \to n$. In other words, the first nodes added have a substantially higher expected degree than those added later. Moreover, the overall power-law degree distribution of the whole graph is a result of the influence of these early nodes.

This correlation between age and degree does not exist for the network of webpages [1]. A brand new webpage can quickly acquire many links and old pages do not have a higher probability of being highly connected. In a reply to this criticism, however, Barabási *et al.* [12] pointed out that this does not mean that preferential attachment does not explain the power-law degree distributions in the world wide web. Rather, the age-correlations imply that the dynamics of the world wide web are more complicated than this simple model: Additional mechanisms, that account for the observed age distribution, could be present.

Furthermore, it is interesting to note that although the growth model contains the non-trivial correlations mentioned above, the correlation-coefficient C approaches zero for $n \to \infty$. Thus, the simple growth model also fails as a viable model for social networks.

1.5 Random Scale Free Networks

The growth model can create networks with scale free degree distributions but, as we have just seen, these networks contain degree-correlations with respect to both neighboring nodes and node-age. If we want to create a truly random network with a scale free degree distribution, how do we go about it?

This question appears innocuous but, as we shall see in the following, the answer is quite interesting. One simple approach is suggested in [72], where the

network is generated by creating n nodes and assigning a number k_i of 'link-stubs' to node i, corresponding to the desired distribution. The network itself is then generated by picking pairs of nodes at random and joining their link-stubs to form complete links. This scheme, however, suffers from one serious problem: It has a tendency to create multiple links between nodes, especially between highly connected nodes. This tendency is easy to understand: Each hub is selected to partner up with other nodes in proportion to its number of links. Thus, two nodes with many links have a high probability of linking to each other many times. Typically this is not the case in real networks[11]. Here, only a single link is allowed to join each pair of vertices.

If we try to impose the condition that only a single link may join two nodes on the stub-connection model, problems arise. The problem is that we always end up with configurations where the remaining stubs have no eligible partners. The authors of [60] performed numerical studies for the asymptotically scale-free network of the Internet [27]. In the resulting network, they found that the average number of unconnected edge stubs is 23 times greater than its standard deviation, precluding even the occasional completion of this algorithm. Thus, this idea is not a viable option.

One promising way of dealing with these problems is introduced in [60]. Again, we begin with n nodes and the desired degree distribution. Then we, create some configuration of the links, where all link stubs can find a partner. This could, for example, be done by first connecting all links from the node with the highest degree to eligible nodes further down the degree hierarchy, then taking the node with the second highest degree, etc. The resulting network is far from random. We can now apply the *local rewiring algorithm* [59] to the network. This algorithm randomizes a network, while strictly conserving degrees of its nodes. The

[11] This statement is not unproblematic, since it assumes that the links of the internet are unweighted. If we take the data regarding the internet on the level of autonomous systems [27] (more on this network in the next section), there exists only one link between the two main hubs (with degrees of 1458 and 750 respectively). The stub-connection model would predict as many as 43 links [60]. It could be argued, however, that the reason there is only one link, is that we are regarding this network as unweighted. Had we weighted this network (for example, by weighting each connection by the real world network throughput), the weight of the link between two hubs would be much higher than most link-weights in the network. In terms of the network adjacency matrix (see equation (2.1)), which is the most common network representation, there is no way to distinguish between multiple links and (integer) weights.

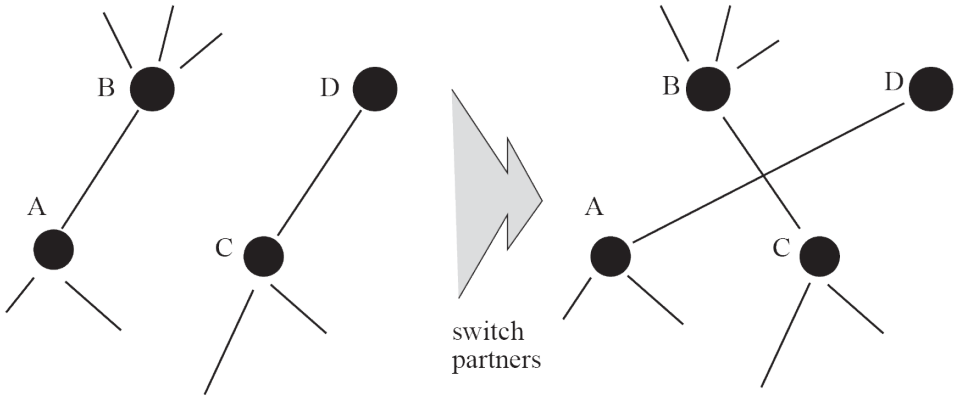


Figure 1.5: One elementary step of the local rewiring algorithm. A pair of links $A \longleftrightarrow B$ and $C \longleftrightarrow D$ is randomly selected. The links are subsequently rewired such that $A \longleftrightarrow D$ and $B \longleftrightarrow C$, provided that none of these edges already exist in the network, in which case the rewiring step is aborted. The last restriction prevents the occurrence of multiple links connecting the same pair of nodes. From [60].

rewiring algorithm consists of repeated application of the elementary rewiring step shown and explained in detail in figure 1.5. It is clear that the number of neighbors of every node in the network remains unchanged after an elementary step of this randomization procedure. The directed network version of this algorithm separately conserves the number of upstream and downstream neighbors (in- and out-degrees) of every node. Randomization of a given network (for example, the non-random network described above) is achieved by repeated application of this rewiring step.

In the case of the random scale free network outlined above, it is interesting to note that the constraint of 'no multiple links' induces an effective repulsion between hubs. This repulsion affects the average degree $\langle k_{\text{neighbor}} \rangle_{k_0}$ of neighbors of nodes with a certain degree k_0. Specifically $\langle k_{\text{neighbor}} \rangle_{k_0} \sim k_0^{-1/2}$ for this network. Precisely this scaling relationship between $\langle k_{\text{neighbor}} \rangle_{k_0}$ and k_0 has been observed for the internet [78]. Thus, we can attribute this relationship to the effective repulsion between hubs. In the simple stub-connection model (where multiple links between nodes are allowed) we have that $\langle k_{\text{neighbor}} \rangle_{k_0} \sim \text{const}$. [72]. As we shall see shortly, this agreement does not necessarily imply that the topology of the random model is similar to the topology of the real internet.

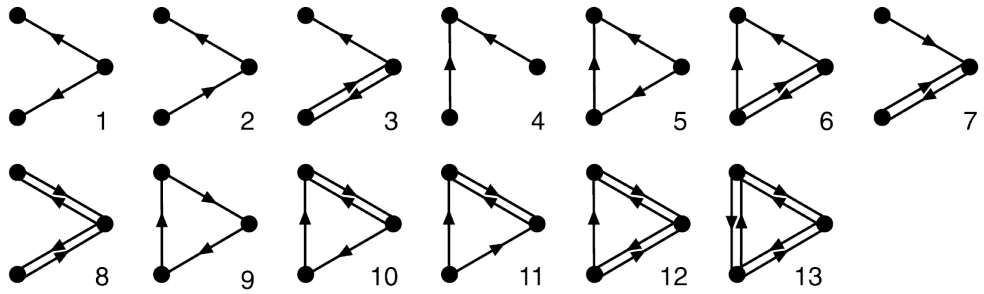


Figure 1.6: Network motifs. Displayed above are the 13 different types of three node connected subgraphs. Motif no. 5 is of particular interest in biological networks; it is called the 'feed forward loop'. From [62].

In summary, the rewiring algorithm can be used to create a random network with any degree distribution. We can also apply this algorithm to any existing network to create a randomized version where the degree distribution is conserved. In fact, we can easily modify this algorithm to conserve almost *any property* that we would like the network to sustain; the generic trick is to only allow rewirings that maintain/emphasize the property we are interested in. In the following, we will make frequent use of this algorithm.

1.6 Motifs: Building Blocks

The clustering coefficient from section 1.2 was a great help in understanding why random networks are not an appropriate model of real world networks. Now that we have the matter of the degree distribution firmly under control, it is useful to return to this *bottom up* approach to understanding networks. In a paper on directed networks, Milo *et al.* [62] analyzed the presence of *network motifs* in a number of different networks. Motifs are sub-graphs of only a few nodes—the triangle used in the definition of the clustering coefficient (see section 1.2) is one example. Figure 1.6 displays all 13 types of three-node connected subgraphs that can occur in directed networks (there are 199 types of motifs of size four). Clearly, directed networks are much richer than undirected networks in this respect; undirected networks have only two motifs for three node

subgraphs and seven four-node motifs.

In order to determine which motifs are important, Milo *et al.* counted the number of occurrences of each motif in a number of real world networks. With this data in hand, the directed version of the local rewiring algorithm from section 1.5 was applied to each network. Comparing the results from the real network with the results from the randomized networks, it is possible to identify the significant motifs; these subgraphs are either suppressed or enhanced compared to the randomized network. Having identified these relevant motifs, we can go back to the original network and begin to interpret the significance of our findings.

Milo *et al.* analyze many different networks. Here, I will focus on one particular finding that regards transcription gene regulation networks, see also [87]. Transcription networks are biochemical networks responsible for regulating the expression of genes in cells. We can think of these as directed networks: Each node represents a gene and each directed link originates from a gene that encodes a transcription factor protein and points towards a gene that is regulated by that transcription factor. The two best characterized such networks are yeast *Saccharomyces cerevieiae*—which is an eukaryote—and the bacterium *Escherichia coli*.

For both of these networks, the feed forward loop (motif no. 5 in figure 1.6) appeared in numbers that are more than 10 standard deviations greater than its mean number of appearances in randomized networks[12]. This important motif arises in the common situation when a transcription factor X regulates a second transcription factor Y, such that both X and Y jointly regulate an operon Z. Feed forward loops are known to have many important functions in regulatory networks [87].

Motifs are the building blocks of networks in the sense that they reflect processes that are typical for the network we are studying. They reflect the functional sub-networks that have merged to form the larger network. At the same time, the motifs can help to shed light on the constraints under which a network has evolved. In this way the motifs shed light on which processes during the network

[12]Other motifs than the feed forward loop appear with a significantly higher frequency in these networks than their randomized counterparts. I mention this motif because of its simplicity and ubiquity in many other regulatory systems.

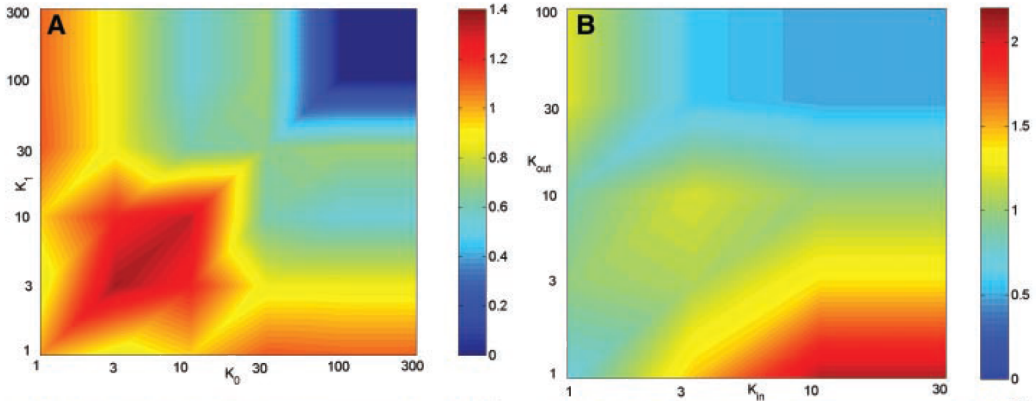


Figure 1.7: Correlation profiles of protein interaction and regulatory networks in yeast. Panel A displays the ratio $R(k_0, k_1) = P(k_0, k_1)/P_r(k_0, k_1)$ for the interaction network, and panel B displays the same quantity for the regulatory network. While the transcription regulatory network is naturally directed, the interaction network in principle has no directionality. Therefore, panel A shows k_0 and k_1 as the total number of neighbors of each of the two nodes, while panel B shows the out- and in-degrees of the two nodes connected by a directed edge $0 \rightarrow 1$. Note that the axes in the two plots are different. From [59].

evolution that have caused their appearance.

1.7 Correlation Profiles

The network motifs focused on a bottom up approach in the study of the structure of networks. Maslov and Sneppen [59] attack the networks from the *top down*. We can recall from section 1.4 that the major problem with the growth models was that although the degree distribution was correct, the model had node-degree correlation that were inconsistent with real world networks. Sneppen and Maslov work from this problem and study the degree-*correlations* of real networks. Specifically, they studied the interaction and transcription regulatory networks in yeast (*Saccharomyces cerevisiae*). The protein interaction network consists of 4549 physical interactions (the proteins are able to bind to each other) between 3278 yeast proteins. The degree distribution of this network has a power-law degree distribution where $P(k) \sim k^{-\alpha}$ with $\alpha = 2.5 \pm 0.3$ in the range $1 < k \leq 100$. We are familiar with the genetic regulatory network from section 1.6. It is formed by 1289 directed (positive or negative) direct tran-

scriptional regulations within a set of 682 proteins.

In order to test for correlations in the node-degrees in each of these two networks, Maslov and Sneppen first calculated the quantity $P(k_0, k_1)$ defined as the normalized probability that a pair of proteins with degrees k_0 and k_1 interact directly with each other on the full set of links. Again, it is especially interesting to compare this with the same quantity $P_r(k_0, k_1)$ calculated on a version of the network randomized with the local rewiring algorithm. Correlations occur where the ratio $R(k_0, k_1) = P(k_0, k_1)/P_r(k_0, k_1)$ is different from unity. In figure 1.7, panel A shows the correlation profile for the undirected interaction network. Here k_0 and k_1 are the total number of neighbors of each of the two nodes, while panel B shows the out- and in-degrees of the two nodes connected by a directed edge $0 \rightarrow 1$ for the directed regulatory network. Thus, $R(k_0, k_1)$ is symmetric for the protein interaction network and not for the regulatory network.

Figure 1.7 reveals the regions on the k_0k_1-plane, where the correlation between node degrees are significantly enhanced or suppressed compared to a randomized network. These types of correlation profiles can supply a great deal of information about the structure of the network we study. Let us continue to investigate the yeast-networks.

For the interaction network, in panel A, along the diagonal, we can observe an enhanced affinity, of proteins with between four and nine interactions, to interact[13]. This feature is explained by a tendency of members of multiproteins to interact with other proteins from the same complex. The red zones in the upper left- and lower right corner reflect the tendency of highly connected nodes to have neighbors of low degree, while the blue area in the upper right corner shows that there is a highly reduced likelihood for two hubs to link to each other.

The regulatory network panel B shows similar patterns. One implication of

[13]Note that for poorly understood reasons, the two-hybrid experimental data have a significant asymmetry between baits and preys, with bait hybrids being more likely to be highly connected than their prey counterparts. This can be seen, e.g., in the fact that average connectivity of baits with at least one interaction partner is close to 3, whereas the same quantity measured for preys is only 1.8. Because each reported interaction involves one bait and one prey protein, this asymmetry needs to be taken into account when constructing an uncorrelated 'null model' for the interaction network.

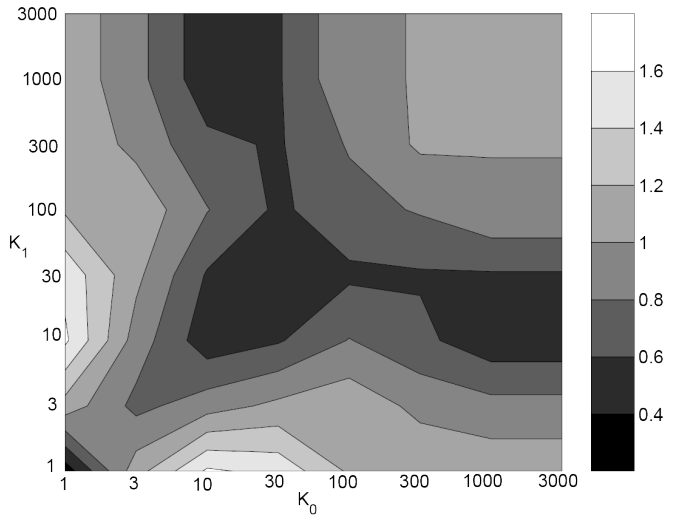


Figure 1.8: Correlation profile $R(k_0, k_1)$ of the internet. We clearly see that $P(k_0, k_1)$ is *not* identical to the randomized version $P_r(k_0, k_1)$ in the case of the internet. See the main text for details. From [60].

the observed structures in both panels of figure 1.7 is the suppression of the propagation of deleterious perturbations over the network. Because highly connected nodes serve as powerful amplifiers for the propagations of any destructive perturbations, it is especially important to limit this propagation beyond the neighbors of these hubs. This property augments the result that we have mentioned earlier (in section 1.3): In general, scale free networks are less susceptible to random attacks but they are more susceptible to attacks on highly connected nodes [7]. The anticorrelation presented by Maslov and Sneppen [59] implies a reduced branching ratio around these nodes and thus provides a certain degree of protection against such attacks.

Later, the same authors [60] analyzed the internet on the (coarse grained) level of hardwired autonomous systems [27]. This network is a snapshot of the internet taken on January 2nd 2000; it consists of 6 474 autonomous systems (nodes) connected by 12 572 links. The correlation profile of this network is displayed in figure 1.8. It is clear from this correlation profile that there are many differences between $P(k_0, k_1)$ and the randomized version $P_r(k_0, k_1)$ of the internet. There is a greatly suppressed probability of links between nodes of low degrees (when $k_0 \leq 3$ and $k_1 \geq 1$). There is a suppression of links between nodes of intermediate degrees (when $k_0 < 100$ and $k_1 \geq 10$), with less suppression as k_0 becomes smaller. Finally there is a pronounced enhancement of the number of links connecting a node of a low degree (approximately when $1 \leq k_0 \leq 3$) to a node of intermediate degree (in the range $10 \leq k_1 \leq 100$). In the upper right

hand corner, we see that $R(k_0, k_1) \approx 1$, which means that the probability of the highest nodes having a connection is approximately the same as in a random network (all highly connected nodes are connected to each other).

Clearly, the correlation profiles in figure 1.7 and figure 1.8 are qualitatively different. The structure of the internet is stratified with a division of nodes into three 'layers' categorized by low, intermediate, and high degrees[14], communicating as described above. The molecular networks are characterized by suppressed connections between nodes of very high degree, and increased number of links between nodes of intermediate degree. Thus, the correlation profile allows one to differentiate between complex networks with similar degree distributions.

1.8 Hierarchies

So far, we have seen the statistical characterization of complex networks move from simple scalar measures, such as the mean and the median, to the distribution of degrees, moving towards more complex measures, such as motifs and the two point correlation profile. A further step in this progression was taken by Trusina *et al.* [91], when they introduced the notion of the *topological hierarchy* in networks. These authors argue that the defining feature of hierarchical systems and organizations is the existence of a hierarchical chain of command. A request starts out at the lowest levels, travels up the ranks and then, after encountering the decision-making level, travels back down to its target. Further, these authors point out that the degree of a node k is a good proxy of the importance of that node. One example is web-pages where the in-degree of a page is a good measure of its popularity; it is easy to think up many other examples where degree corresponds to 'importance'.

Thus, it is natural to define a *hierarchical path* between two nodes in a network as consisting of an *up path*, where one is allowed to step from node i to node j only if their degrees k_i and k_j satisfy $k_i \leq k_j$, followed by a *down path* where only steps to nodes of lower or equal degree are allowed. In a hierarchical path, either the up- or down path is allowed to have zero length. It is possible to calculate the shortest paths between all pairs of nodes in a network and calculate the fraction

[14] This may be due to the stratified structure of actual internet into users, low-level (possibly regional) internet service providers (ISP), and high-level (global) ISP.

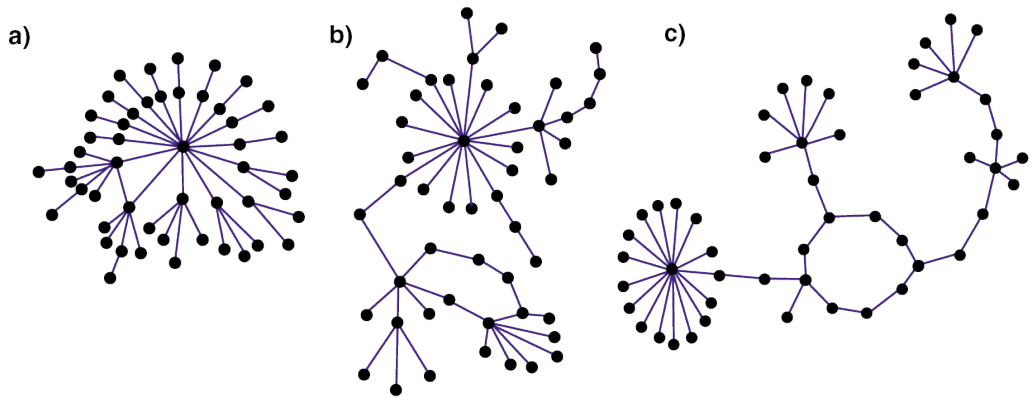


Figure 1.9: Examples of networks. Displayed here are three networks with $n = 50$ nodes and degree distribution $p(k) \sim k^{-\gamma}$ with $\gamma = 2.5$. Panel a) displays the maximally hierarchical version, b) is the randomized version and panel c) is the maximally anti-hierarchical version. From [91].

of these that are hierarchical paths. This fraction of all nodes is denoted $\mathscr{F}$.

We can also imagine pairs of nodes, where the shortest path is non-hierarchical (a non-hierarchical *s*hortcut) but where a hierarchical path (longer than the shortest path) between the two nodes does exist; this fraction of all shortest paths is denoted $\mathscr{S}$; and the fraction pairs of nodes, where no hierarchical paths exists between them is denoted $\mathscr{U} = 1 - \mathscr{F} - \mathscr{S}$. For any given network, it is instructive to compare these three quantities with their values calculated on a network randomized by the local rewiring algorithm. As was the case in the previous sections, this comparison will allow us to spot design in networks; to see where the network possesses structure that does not stem from randomness. We call the counterparts calculated on randomized networks $\mathscr{F}_r, \mathscr{S}_r$ and $\mathscr{U}_r$.

With these measures of hierarchy in hand, it is natural to analyze the hierarchical structure of several real world networks. Trusina *et al.* have explored the following networks; note that all of these are undirected.

(i) **The internet on the level of hardwired autonomous systems** (from [27]). This network of 6 474 nodes and 12 572 links was analyzed in [60] and discussed in section 1.7. In this network, the level of topological hierarchy is high, with $\mathscr{F} = 0.95$. So is the level of hierarchy in its randomized counterpart,

with $\mathscr{F}_r = 0.99$. The small reduction in $\mathscr{F}$ compared to the randomized version is mainly due to a higher number of non-hierarchical shortcuts in the real internet, with $\mathscr{S} = 0.02$ compared to $\mathscr{S}_r = 0$.

(ii) **The largest component of the yeast protein interaction network.** This network with 2 839 nodes and 4 220 links is part of one of the two networks analyzed in [59] and discussed in section 1.7. As we anticipate from the strong preference for hubs to connect to nodes with low degree and to avoid connections to other highly connected nodes seen in figure 1.7, this network is highly anti-hierarchical with $\mathscr{F} = 0.33$. The randomized version of this network has $\mathscr{F}_r = 0.88$. This is yet another confirmation of a structure where highly connected nodes are placed on the periphery of the network. As was the case with the internet, the randomized version of the network has a much lower fraction of non-hierarchical shortcuts, $\mathscr{S}_r = 0.02$, compared to the actual data, where $\mathscr{S} = 0.17$. One possible explanation for this phenomenon, both in the internet and in the protein interaction network, is a natural tendency toward shorter distances and thus shorter and more specific signaling.

(iii) **A scale free email network of correspondence from Kiel University** (from [24]). The largest connected component with 25 151 nodes and 199 963 links compiled over a 112 day period. Similarly to the internet, the randomized version is a little more hierarchical than the real network, with $\mathscr{F} = 0.97$ and $\mathscr{F}_r = 0.98$.

(iv) **The network of CEOs.** This network consists of 6 193 Executive Company Directors linked by 43 077 memberships of boards of directors. As could be expected, this network is less hierarchical than the internet and email networks, but much more hierarchical than the protein interaction network. Specifically we have $\mathscr{F} = 0.78$ and $\mathscr{F}_r = 0.84$.

The notion of topological hierarchy has supplied us with a powerful tool to understand and quantify the structural differences between the internet and the protein interaction network that we began to study using the correlation profile.

The rewiring process (from section 1.5) used to create the randomized versions of the networks can be subtly altered to help create *maximally hierarchical* or *maximally anti-hierarchical* versions of a given network. For example, for the

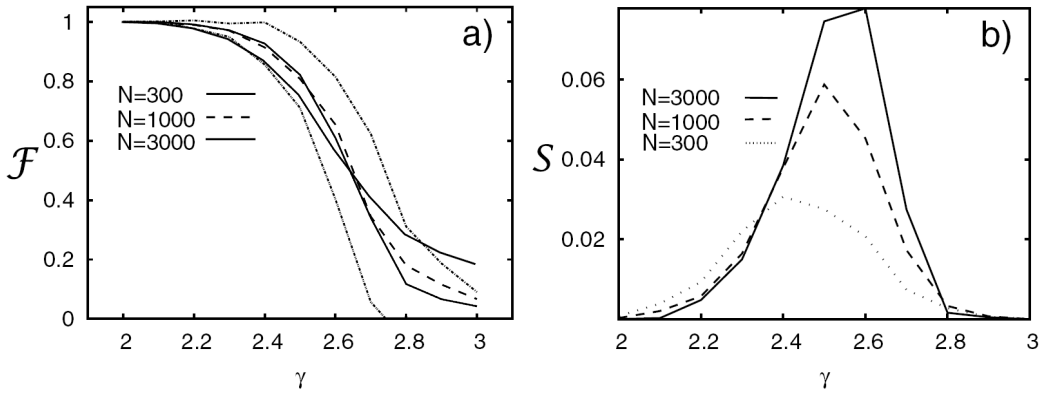


Figure 1.10: Measures of hierarchy vs. the power-law exponent, γ, for random scale free networks. Panel a) displays the fraction of hierarchical paths, $\mathscr{F}$, measured as a function of γ and panel b) shows the fraction of non-hierarchical shortcuts, $\mathscr{S}$, as a function of γ. Both panels display the data for three system sizes $n = 300, 1000, 3000$. From [91]

maximally hierarchical case, one only need to add is a particular preference for reconnection. At each step, we select two links and connect the node with the highest degree to the node with the next highest degree. The remaining two nodes are then linked. Again, multiple links are forbidden and the network must remain connected. To create the maximally anti-hierarchical network, the same procedure is followed, but in this case we connect the node with highest degree to the node with the *lowest* degree and then connect the two remaining nodes. These procedures have been used to generate the different networks in figure 1.9. Applying the algorithm to each of the four test networks it is possible to achieve the limits where $\mathscr{F} = 1$ for maximal hierarchy and $\mathscr{F} \approx 0$ for maximal anti-hierarchy.

Finally, these authors demonstrate that for random scale-free networks designed using the local rewiring algorithm there exists simple correlations between the slope of the asymptotic power law γ and the measures of topological hierarchy $\mathscr{F}$ and $\mathscr{S}$ [91]. Figure 1.10 a) shows $\mathscr{F}$ plotted against γ. We can observe a smooth transition from $\mathscr{F} = 1$ for $\gamma = 2$ to $\mathscr{F} \approx 0$ for $\gamma = 3$. The transition point is weakly dependent on system size and occurs at $\gamma \approx 2.6$. This result is augmented by the plot of $\mathscr{S}$ vs. γ in figure 1.10 b). Here, $\mathscr{S} = 0$ for both $\gamma = 2$ and $\gamma = 3$ but grows smoothly to a maximum at an intermediate value after which it falls off again.

Let start investigating the behavior in these plots. When $\gamma = 2$ we have $\mathscr{F} = 1$ and $\mathscr{S} = 0$. In this state, the tail of the distribution is particularly heavy (recall from equation 1.6 that the mean is not defined for this distribution) and there are many highly connected hubs; the average distance in this network approaches 2, since the majority of nodes are connected via at least one hub. As a result, most shortest paths are via a hub and therefore the network is maximally hierarchical with $\mathscr{F} = 1$ and $\mathscr{S} = 0$. When $\gamma = 3$ the topology of the network is very tree-like. This means the number of alternative paths approaches zero so, again $\mathscr{S} \to 0$. In the intermediate regime $\mathscr{S}$ is not constrained by the two factors that drive $\mathscr{S}$ to zero.

The reason $\mathscr{F} \to 0$ when $\gamma \to 3$ is more complex and the details will not be discussed here. Qualitatively, the reason is that when $2 < \gamma < 3$, nodes with a low degree have a low probability of neighboring a node with higher degree than itself. At the same time, $P(k_{\text{neighbor}} > k) \to 1$ for increasing k, resulting in a hierarchical core of highly connected nodes. A consequence of the combination of these two effects is that many nodes of low degree 'escape' the hierarchy, causing a high fraction of the network paths to become non-hierarchical. For values of γ higher than three, $P(k_{\text{neighbor}} > k) \to 0$ with the degree; thus for these high values of γ the network becomes modular with each of the modules centered around a local hub.

I would like to emphasize that the results above *only* apply to randomized networks. Real networks can—as we have just seen in the examples—arrange their structure quite differently than what is dictated for their randomized counterparts.

CHAPTER 2

Communities

COMMUNITY structure can be defined as the tendency of nodes in some networks to be organized in modules with a high density of links between the nodes inside of a given module and few links between the different modules. Many of the measures considered in the previous section cannot detect such community structure.

Consider a network with C communities of roughly the same size and statistical properties, linked by a small number of inter-community links. The mean, median, etc., provide no signs to suggest that the network possesses community structure. The degree distribution might yield a subtle hint: One could imagine that if a network consists of many small communities, the degree distribution would have a cut-off around the size of the largest community—simply because the hubs would not be able to link to a large number of nodes outside their own community without blurring the community structure. Many other explanations for a cut-off of the degree distribution could, however, exist.

Now, consider the higher order measures. We can argue that, as long as motifs are small compared to the typical community size, counting the relative occurrence of different motifs is un-informative with respect to community struc-

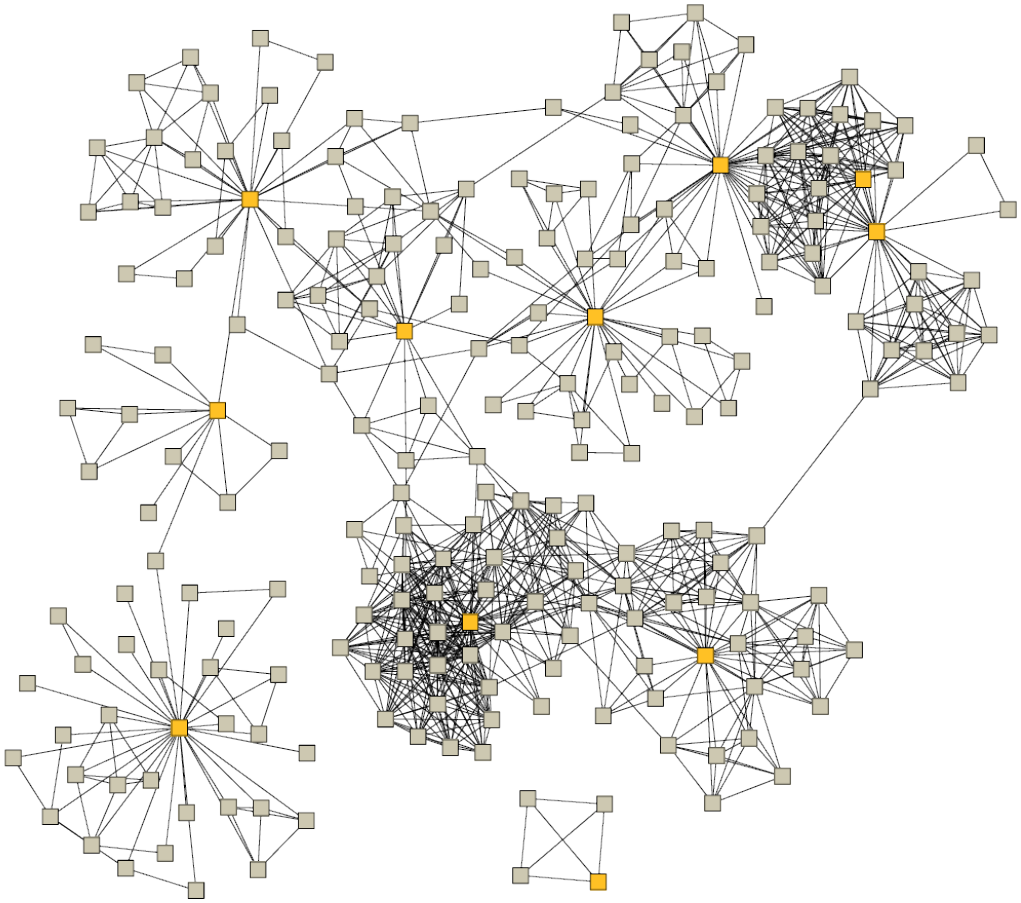


Figure 2.1: A small social network with community structure. Note that the various communities are quite heterogeneous. From [70].

ture. There is no reason to suspect that the global count of motifs is materially changed due to the systematic variation of link-density that communities entail[1]. One would suspect that community structure would somehow manifest itself in the correlation profiles but, in analogy with case of the degree distribution, all such systematic patterns could stem from many other causes. Thus, the correla-

[1]The motifs, however, might help us pinpoint the inter-community *links*: Links between communities rarely participate in triangles. Therefore, one approach would be to count the number of triangles that each link belongs to; the links that belong to an unusually low number of triangles are prime candidates for inter-community links, cf. [81]. Identifying the inter-community links will allow us detect the communities. A different use of motifs, is seen in section 2.4, where specific classes of motifs, called k-cliques, are utilized to *define* communities.

tion profile does not provide a measure that enables one to derive the existence of community structure. The hierarchy measures do allow us to detect some signs of communities. In a network with community structure, the number of hierarchical paths is low: If a network consists of a number of nearly disjunct hierarchical communities, a large fraction of the shortest paths across communities would have to move up and down the degree hierarchy making these paths non-hierarchical.

The situation, however, is more problematic that we might expect. Consider figure 2.1 and note how the different communities are heterogeneous; they have varying sizes and each community has different link densities and degree distributions. This situation does not make it any easier for the measures from chapter 1 to detect the community structure. In fact, if the communities are heterogeneous, the global measures can be directly misleading! Several real world networks possess such heterogeneous communities. One such network is the networks of sexual contact, where separate communities of high- and low-activity individuals have been observed [29]. A characterization of this network by a single figure, for example the median number of partners, would result in misleading descriptions of features of the network. These problems thus are directly relevant to important subjects, such as epidemiological dynamics [35].

In summary, if we want to understand the structure of networks, it is of paramount importance to understand how to identify communities. In the context of the 'history of structure', the work on community detection has developed quite independently of the work described in chapter 1. A full merger of these two lines of research has not yet taken place. Until we discover a way to unify these different paths, the ambition of a coherent and comprehensive set of statistical tools to understand networks will remain elusive. For now, a general strategy is clear. When we have determined the modular structure of a given network, we can proceed to analyze these modules using the tools of the previous chapter. The remainder of this chapter is devoted to the community structure of complex networks. Other reviews of community detection, in complex networks, can be found in [20, 66].

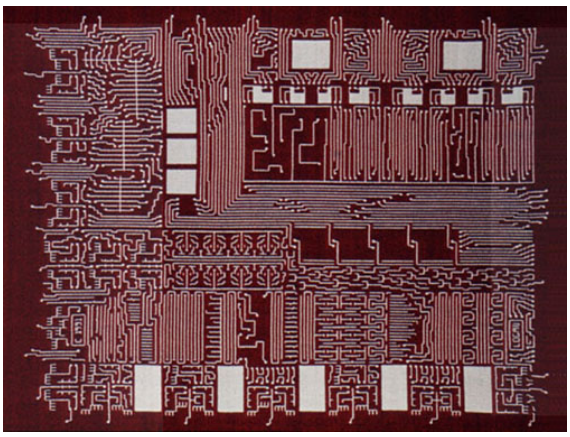


Figure 2.2: A microchip. In computer science, graph partitioning has been used in computers with multiple processors in order to place processes (nodes) which need to communicate (links) on the same processor. The fastest configuration has the fewest cross-processor links. The chip displayed here is actually a work of art entitled *Microchip Series 2:A* (1991) weaved on wool by the artist W. Logan Fry. (The actual size is 45" × 35". It is part of the collection of The Minneapolis Institute of Arts.)

2.1 Spectral Bisection

Dividing a network into C modules with as few links between the modules as possible, has been of interest to computer scientists for many years. This is because graph partitioning facilitates faster parallel processing. Consider distributing n computational tasks onto C processors. Further, imagine that there are m links between these tasks, each link signifying that the two tasks need to communicate. The crucial point is that because communication between two tasks running on the same processor is fast and cross-processor communication is slow, we would like to distribute the tasks in such a way that *the number of cross-processor links is minimized*. Minimizing this number of links is identical to the graph-partitioning problem on a graph with n nodes, m links and predetermined community sizes. The minimal number of links is called the *cut-size* and denoted by R. In the context of complex network structure, we are interested in a slightly different problem. We wish to *detect* naturally occurring community structure, and therefore we have no pre-determined idea of the community sizes. We would like the algorithm to tell us what size the communities are.

Minimization of the cut-size is an integer programming problem and can be solved exactly in polynomial time [32]. The leading order, however, of the polynomial in question is n^{C^2}, which is prohibitively slow even dividing into $C = 2$ communities. Fortunately, many approximate methods exists; in this section, we will discuss a particularly popular method, called *spectral bisection*,

which was developed by Fiedler [26] and popularized by Pothen *et al.* [80]. I will discuss the details of this method below.

First, we phrase the problem more precisely. Consider an unweighted and undirected graph with n nodes and m links. This network can be represented by an *adjacency matrix* $\mathbf{A}$, with elements

$$A_{ij} = \begin{cases} 1, & \text{if there is a link joining nodes } i \text{ and } j \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

This matrix is symmetric with $2m$ entries[2]. Connecting this representation of the network to the concepts from the previous chapter, we note that the degree of node i can be found by summing over the adjacency matrix

$$k_i = \sum_j A_{ij} \,. \tag{2.2}$$

The cut-size can be expressed in terms of $\mathbf{A}$. We find that

$$R = \frac{1}{2} \sum_{ij} [1 - \delta(c_i, c_j)] A_{ij}, \tag{2.3}$$

where c_i is the community to which node i belongs and $\delta(\alpha, \beta) = 1$, if $\alpha = \beta$ and $\delta(\alpha, \beta) = 0$, if $\alpha \neq \beta$. We wish to minimize R.

To simplify the problem as much as possible (a more complete solution is presented in chapter 6), we limit ourselves to the case where $C = 2$ (hence, 'bisection'). Let $\mathbf{z}$ be a vector (called the *partition vector*) with n elements and component $z_i = 1$, if node $i \in c_1$, and $z_i = -1$, if node $i \in c_2$. We can utilize $\mathbf{z}$ to emulate our delta function, by realizing that

$$z_i z_j = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are in the same community} \\ -1, & \text{if } i \text{ and } j \text{ are in different communities.} \end{cases} \tag{2.4}$$

Further, this implies a normalization $\mathbf{z}^T \mathbf{z} = n$. Using the partition vector, we express the cut-size as

$$R = \frac{1}{4} \sum_{ij} [1 - z_i z_j] A_{ij} \,. \tag{2.5}$$

[2]If the network had been directed, the adjacency matrix would have been asymmetrical, and, if the network had been weighted, link weights would have replaced the 1's, in equation (2.1).

We can convert this to a matrix equation. First, we note that the first term in equation (2.5) can be written as

$$\sum_{ij} A_{ij} = \sum_i k_i = \sum_i z_i^2 k_i = \sum_{ij} z_i z_j k_i \delta(i,j), \tag{2.6}$$

where we have used the definition of $\mathbf{z}$. The entire expression condenses to

$$R = \frac{1}{4} \sum_{ij} z_i z_j (k_i \delta(i,j) - A_{ij}) = \frac{1}{4} \mathbf{z}^T \mathbf{L} \mathbf{z}, \tag{2.7}$$

where $\mathbf{L}$ is the so called *Laplacian* matrix, defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{2.8}$$

where $\mathbf{D}$ is a diagonal matrix with $L_{ii} = k_i$. In simple terms, our task is to find the $\mathbf{z}$ that minimizes the cut-size. As we shall see shortly, this is closely related to the spectrum of the Laplacian.

Therefore, let us take a closer look at this matrix. First of all, we note that by construction, the sum of each row (and column) of $\mathbf{L}$, is equal to zero. This implies that the vector $c\,\mathbf{e}$, where $\mathbf{e} = (1,1,\ldots,1)^T$, is always an eigenvector of $\mathbf{L}$ with eigenvalue zero. Further, the Laplacian is positive semi-definite. A matrix $\mathbf{M}$ is positive semi-definite when

$$\mathbf{x}^T \mathbf{M} \mathbf{x} \geq 0 \quad \text{for all } \mathbf{x} \in R^n, \tag{2.9}$$

which implies that all eigenvalues of $\mathbf{M}$ are real and non-negative. The Laplacian, $\mathbf{L}$, fulfills Eq (2.9). Let us show this. First, note that we can think of the Laplacian as a sum of matrices, each of which correspond to one link between two vertices i, j. Such a 'mini-Laplacian' has zeros everywhere, except in four positions:

$$\mathbf{L}_{i \leftrightarrow j} = \begin{bmatrix} \ddots & & & & \\ & 1_{ij} & & -1_{ji} & \\ & & \ddots & & \\ & -1_{ji} & & 1_{ij} & \\ & & & & \ddots \end{bmatrix}, \tag{2.10}$$

where the notation $i \leftrightarrow j$ means that there is a link between node i and j, and 1_{ij} means that there is a '1' in position (i,j). We note that

$$\mathbf{x}^T \mathbf{L}_{i \leftrightarrow j} \mathbf{x} = (x_i - x_j)^2. \tag{2.11}$$

Next, we simply insert the Laplacian in Eqn. (2.9)

$$\mathbf{x}^T\mathbf{L}\mathbf{x} = \mathbf{x}^T\left(\sum_{\text{links}}\mathbf{L}_{i\leftrightarrow j}\right)\mathbf{x} \tag{2.12}$$

$$= \sum_{\text{links}}(x_i - x_j)^2. \tag{2.13}$$

Clearly, this sum must be greater than or equal to zero and thus, the Laplacian is positive semi-definite. Without loss of generality, we can arrange eigenvalues in increasing order $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. The eigenvalue $\lambda_1 = 0$ is always the smallest eigenvalue of the Laplacian and a corresponding eigenvector is $\mathbf{v}_1 = \mathbf{e}/\sqrt{n}$, normalized to unit length[3]. Further, note that, due to orthogonality, all other eigenvectors must have both positive and negative elements.

Now, let us write the partition vector $\mathbf{z}$ as a linear combination of normalized eigenvectors $\mathbf{v}_i$ of $\mathbf{L}$. In other words

$$\mathbf{z} = \sum_{i=1}^{n} a_i\mathbf{v}_i\,, \tag{2.14}$$

where $a_i = \mathbf{v}_i^T\mathbf{z}$. The normalization of $\mathbf{z}$ gives us

$$n = \mathbf{z}^T\mathbf{z} = \sum_i a_i^2\,. \tag{2.15}$$

Inserting equation (2.14) into R from equation (2.7), we find

$$R = \frac{1}{4}\sum_i a_i\mathbf{v}_i^T\mathbf{L}\sum_i a_i\mathbf{v}_i \tag{2.16}$$

$$= \frac{1}{4}\sum_{ij} a_i a_j\lambda_k\delta(i,j) \tag{2.17}$$

$$= \frac{1}{4}\sum_i a_i^2\lambda_i\,, \tag{2.18}$$

where λ_i is the eigenvalue of $\mathbf{L}$ that corresponds to the eigenvector $\mathbf{v}_i$, and we have made use of the orthonormality of the eigenvectors. With eigenvalues arranged in increasing order, minimizing R is a question of the following: We must

[3] If the graph contains disjoint components and thus breaks perfectly into communities with no joining links, this eigenvalue is degenerate. To see this, note that a simple permutation of rows and columns can make the Laplacian block diagonal. Each diagonal block will form the Laplacian of its own component and, also, have an eigenvector $\mathbf{v}_k$ with eigenvalue zero.

select a_i^2, such that, in the sum of equation (2.18), as much weight as possible is put in the terms corresponding to the lowest eigenvalues, and as little weight as possible is placed in the terms corresponding to the highest eigenvalues—while respecting the normalization constraint, equation (2.15).

The partition that minimizes R arises when $\mathbf{z} \sim \mathbf{v}_1$. Then all weight in equation (2.18) is placed in the term corresponding to the $\lambda_1 = 0$ eigenvalue and all other terms are zero since $\mathbf{e}$ is an eigenvector of $\mathbf{L}$ and, therefore, orthogonal to the remaining eigenvectors. Physically, this means that we get the smallest cut-size ($R = 0$) if we choose c_1 as a community that includes the entire graph and c_2 to be completely empty. This solution is technically correct but also quite uninteresting.

A more interesting solution to the problem can be forced by introducing the constraint on the community sizes n_1 and n_2. As we mentioned earlier, this is a natural constraint from the perspective of computer science because, typically, a fixed number of tasks can run on each chip in a computer with multiple processors. In the context of network community detection, this constraint is less natural because we would like to learn about size of the communities. More on this subject later.

Fixing the community sizes also fixes the coefficient a_1^2, since

$$a_1^2 = (\mathbf{v}_1^T \mathbf{z})^2 = \frac{(n_1 - n_2)^2}{n}. \tag{2.19}$$

Naively, we would think that R would be minimized by choosing $\mathbf{z}$ proportional to the second eigenvector[4], of the Laplacian $\mathbf{v}_2$. This puts all weight in equation (2.18) in the term corresponding to the second smallest eigenvalue λ_2 (the *Fiedler value*) and, again, all other terms are zero, due to orthogonality.

The problem is, however, that while the values of the partition vector $\mathbf{z}$ are restricted to ± 1, the elements of $\mathbf{v}_2$ are real numbers, constrained only by the orthonormalization conditions. This means that we cannot choose $\mathbf{z}$ parallel to $\mathbf{v}_2$. In practice, good approximations occur when $\mathbf{z}$ is chosen, such that it is as parallel to $\mathbf{v}_2$ as possible. The two vectors are most parallel, when the product $\mathbf{v}_2^T \mathbf{z}$ is maximal. Since all elements of $\mathbf{z}$ are ± 1, this maximum occurs when the

[4]The fact that this vector has its own name (it is called the *Fiedler*-vector) would appear to indicate that we are on the right track.

ith element of $\mathbf{z}$ has the same sign as the ith element of $\mathbf{v}_2$.

The fixed community sizes n_1 and n_2 may not correspond to the number of positive/negative elements of the Fiedler-vector. If this is the case, the best solution is achieved, when we order the elements of the Fiedler vector from most positive to most negative and assign vertices to one of the groups in the order of these elements, until the groups have the required sizes. If $n_1 \neq n_2$, we can do this in two different ways, one in which the smaller group corresponds to the most positive elements of the vector and one in which the larger group does. We can choose between them by calculating the cut size for both cases and keeping the one that yields the best result.

Since we do not know the sizes of the communities prior to our analysis of the network, it is tempting to disallow the solution $\mathbf{z} \sim \mathbf{v}_1$ by use of an axiom (we could, for example, denote this the 'trivial' solution) and let the community sizes be determined by the sign of the elements of the Fiedler-vector. Apart from a certain amount of philosophical discomfort from the introduction of an *ad hoc* axiom, this approach is problematic because such solutions tend to favor small groups of nodes that are connected to the network by only a few links. In order to remedy this problem, many variations of the cut size criterion have been suggested, for example

- **Gap Cut.** The network is divided according to the largest gap (numerical difference) between two adjacent elements in the list of ordered Fiedler-vector components [90].
- **Ratio Cut.** The ratio cut proposes that the quantity $R/(n_1 n_2)$ should be minimized [95]. This forces the two communities to have roughly the same size.
- **Normalized Cut** The normalized cut consists in optimizing $R/a_1 + R/a_2$, where a_i designates the total number of links from nodes in c_i to the total network [88].

A large body of literature regarding the optimization of these and other criteria exists and the reader is referred to [33], for a review. Some of these methods work quite well for many general network partitioning problems.

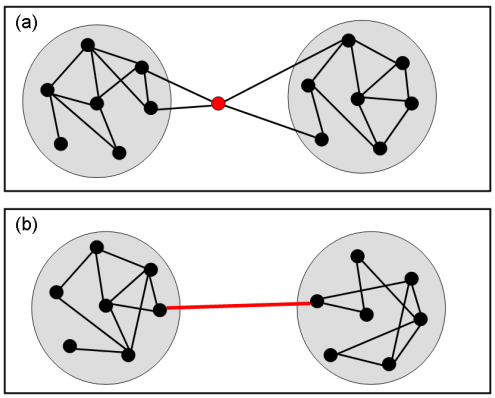


Figure 2.3: Betweenness centrality and edge betweenness. Panel (a) shows the idea of betweenness centrality. The red node in the center of the illustration has the highest betweenness in this network because all shortest paths from one community to the other must pass through this node. Therefore, this node is central to communication in this network. Panel (b) illustrates edge betweenness. The red link has the highest edge centrality in the network because it lies between the two communities.

2.2 From Betweenness to Modularity

One of the first algorithms for detection of communities to emerge from within the complex network community was the *betweenness algorithm* by Girvan and Newman [31,74]. The algorithm is based on a concept called *edge*[5] *betweenness*. Edge betweenness is inspired by the concept of *betweenness centrality*, known from sociology [30]. For a given node its betweenness centrality is the number of shortest paths that pass through the node. The basic idea behind betweenness centrality is that nodes that occur on many shortest paths between other nodes, are central to a network. Betweenness centrality is illustrated in figure 2.3 (a); the red node has the highest betweenness centrality. This is due to the fact that all shortest paths running between the two communities must pass through this node. In this sense the node is central to the network.

Similarly, the edge betweenness of a link is defined as the number of shortest paths that runs through it. The concept of edge betweenness is displayed in figure 2.3 (b). The red link has the highest edge betweenness because all shortest paths between nodes in the two communities must run through this link. It is clear from this definition that the links, that lie between communities, will have higher edge betweenness than other links. In practice the edge betweenness is determined simply by finding the shortest path between all pairs of vertices on a graph and then counting the number of these shortest paths that run along each

[5]Somtimes the words 'vertex' and 'edge' are used to describe 'node' and 'link', respectively. In mathematical graph theory, the former are the standard terms.

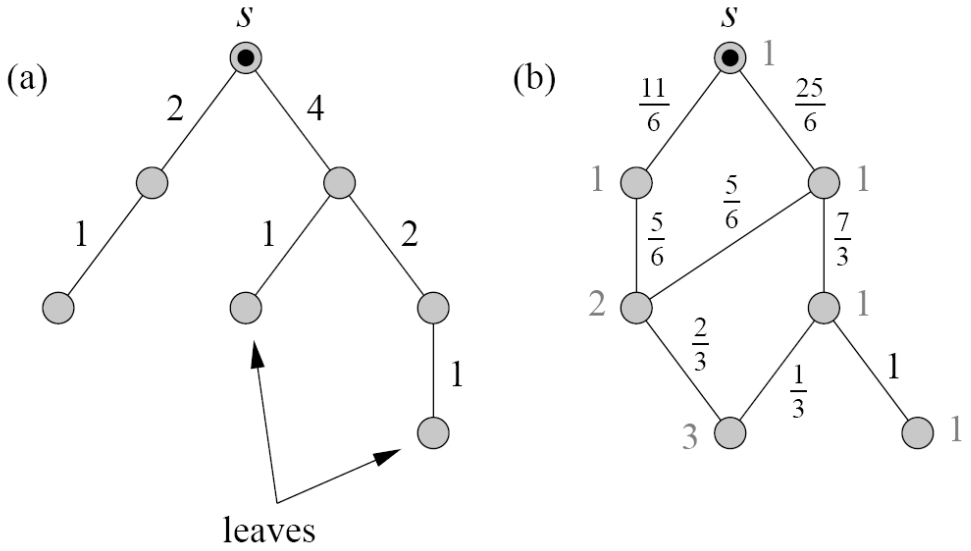


Figure 2.4: Calculation of betweenness: (a) When there is only a single shortest path from a source node s (top) to all other reachable nodes, those paths necessarily form a tree, making the calculation of the contribution to betweenness, from this set of paths, particularly simple. (b) For cases with more than one shortest path to some nodes, the calculation is more complex. First, we must calculate the number of paths from the source to each other node (numbers on nodes). Then these are used to weight the path counts appropriately. In either case, we can check the results by confirming that the sum of the betweennesses of the edges connected to the source node s, is equal to the total number of reachable nodes: six in each of the cases illustrated here. From [74].

edge. For a concrete example, see figure 2.4.

The algorithm Girvan and Newman suggested, uses the edge betweenness as an indicator of where to divide the network. In practice they iteratively remove the link with highest betweenness. The general form of the algorithm becomes:

1. Calculate betweenness scores for all edges in the network.
2. Find the edge with the highest score and remove it from the network.
3. Recalculate betweenness for all remaining edges.
4. Repeat from step 2.

What especially separates this algorithm from earlier attempts to create algorithms that divide networks in to communities by gradually removing nodes, is the recalculation step. This turns out to be crucial since the distribution of betweennesses can change radically when just one edge is removed. Newman and Girvan show, using specific examples on artificial data, that this is indeed the case.

The betweenness algorithm is known to perform very well on many networks. As nodes are removed, the algorithm slowly breaks the network into components which are hierarchically connected through a dendrogram. This dendrogram ends in individual nodes. A partition where each community is its own community is clearly too fine a division. The \$64 000 question is: "When do we stop the link removal?" In order to answer this question, Newman and Girvan introduce a quantity called the *Modularity*, Q, that measures how well a network breaks into communities—i.e. how modular it is [74]. The modularity is inspired by earlier work by Newman [64].

Girvan and Newman argue that networks are modular—not when there are few edges between communities (as was expressed by the cut-size)—but rather when there are fewer edges than expected. Schematically, a function that measures the goodness of a split into communities should therefore be defined by

$$\begin{aligned} Q &= \text{fraction of edges within communities} \\ &- \text{expected fraction of such edges}. \end{aligned} \tag{2.20}$$

To express this quantitatively, we create a $C \times C$ symmetric matrix, $\mathbf{E}$, where element e_{ij} denotes the fraction of the total number of links that run from nodes in community i to nodes in the community j.

The trace of $\mathbf{E}$ yields the fraction of all links that are intra-community links. This sum corresponds to the positive term in equation (2.20). A good division into communities will have a trace that is high. A trace close to unity, however, is not necessarily a good indicator of community structure on its own; a trace of one could be obtained simply by placing all nodes in one community. The sum across this matrix $a_i = \sum_j e_{ij}$ represents the fraction of all links that connect to nodes in community i. In a network, where links fall without regard for community structure, we would expect the fraction $e_{ij} = a_i a_j$. Thus, this corresponds to the negative term in equation (2.20). We can now formulate

equation (2.20) precisely

$$Q = \sum_{i=1}^{C} (e_{ii} - a_i^2). \tag{2.21}$$

If the nodes are placed in communities at random, we get $Q = 0$ and for values of $Q = 1$, which is the maximum, we have good community divisions.

Now, the betweenness algorithm is complete. For each split of the network until we reach the node level, we simply calculate Q. Then, we go back and pick the division of the network for which Q is maximal. In practice this algorithm works very well. It does, however suffer from one detrimental flaw; computation of all shortest paths in each step is very expensive. The complexity of the algorithm scales with the number of nodes n and number of edges m as $O(m^2 n)$. Thus this algorithm is only applicable to rather small graphs with an approximate upper network size limit at around 10^5 nodes.

The goal of the betweenness algorithm is to find high values of Q, but the algorithm is very slow. One is tempted to ask the question "Why not simply optimize Q?"—this might be much faster than running the costly betweenness algorithm. Newman realized this option in two later papers [18, 73], where he maximized Q on very large networks using a simple greedy optimization.

By doing this, Newman started a whole new sub-field of community detection, dedicated to the understanding of general properties of Q and discovering ways of optimizing Q on various networks. I will not spend time on strategies for Q-optimization here, since the subject is treated in detail in chapter 6. Relevant papers in this tradition include [8, 34, 68, 69].

It is, however, of interest here, to note a connection between the modularity and the cut-size from Eq (2.3). To see this connection explicitly, we must rewrite Q in terms of the adjacency matrix. We begin by fleshing out what is meant by the elements of the **E** matrix, in terms of the adjacency matrix: Here,

$$e_{ij} = \frac{1}{2m} \sum_{vw} A_{vw} \delta(c_v, i) \delta(c_w, j), \tag{2.22}$$

which is still the fraction of total links that join nodes in community i with

nodes in community j. And the row sum a_i is

$$a_i = \frac{1}{2m}\sum_v k_v \delta(c_v, i), \tag{2.23}$$

which is again the fraction of all links that are attached to nodes in community i. We insert this in equation (2.21)

$$\begin{aligned} Q &= \sum_i (e_i - a_i^2) && (2.24)\\ &= \sum_i \left[\frac{1}{2m}\sum_{vw} A_{vw}\delta(c_v,i)\delta(c_w,i) \right. \\ &\qquad \left. - \frac{1}{2m}\sum_v k_v \delta(c_v,i)\frac{1}{2m}\sum_w k_w \delta(c_w,i) \right] && (2.25)\\ &= \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\sum_i \delta(c_v,i)\delta(c_w,i) && (2.26)\\ &= \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\delta(c_v,c_w). && (2.27) \end{aligned}$$

The final expression displays the similarities with the cut-size. The modularity maximizes the number of intra-community links rather than minimizing the number of inter-community links (this is the difference between the multiplication by $[1-\delta(c_i,c_j)]$ in the case of the cut-size, and the multiplication by $\delta(c_i,c_j)$ seen here).

The extra term $(k_i k_j)/(2m)$ can be interpreted as a 'null model' containing our idea of what to expect of the network (except for the community structure). We see explicitly that this term makes the assumption that (if there are no communities) the probability of a link between two nodes is proportional to the product of the two degrees. This is a sound assumption, since the degree distribution is an important feature in most networks. By the normalization $(2m)^{-1}$, this term incorporates information about the entire network into the modularity. The cut-size has no equivalent term. The similarities and differences between cut-size and modularity will be illuminated further, in the following section.

Finally, we note that the formulation of Q in Eq (2.27) allows us to maximize it using many of the tools that are already available for the cut-size (with small adjustments). In [69], for example, a method of spectral optimization of Q is

discussed; the calculations there are analogous to the spectral optimization of R described in section 2.1.

2.3 Communities and Spin States

In a recent paper [83], Reichardt and Bornholdt have elegantly incorporated the primary idea behind the modularity and the ratio cut, into a very general framework. In spectral bisection, we minimized the number of links between communities, in the case of modularity, we maximize the number of intra-community links (modulo the null model). Both of these measures yield good results, but the form of the modularity, especially, can appear quite *ad hoc*.

The idea is to systematically design the most general measure that optimizes the number of links and non-links both inside- and between communities; this will result in a framework that we can utilize to gain a deeper understand the other measures. There are four possibilities:

(i) First, we can reward internal edges between nodes of the same community. (This relates to modularity.)

(ii) We can also penalize edges that are *not* present between nodes of the same community.

(iii) Following this idea, it is natural to penalize edges between existing groups. (This relates to cut-size.)

(iv) Finally, we can reward non-links between nodes of different communities.

Combining all of these into one single criterion H that we would like to minimize, gives us

$$\begin{aligned} H = & - \sum_{ij} \alpha_{ij} A_{ij} \delta(c_i, c_j) && (2.28) \\ & + \sum_{ij} \beta_{ij} (1 - A_{ij}) \delta(c_i, c_j) && (2.29) \\ & + \sum_{ij} \gamma_{ij} A_{ij} (1 - \delta(c_i, c_j)) && (2.30) \\ & - \sum_{ij} \eta_{ij} (1 - A_{ij})(1 - \delta(c_i, c_j)), && (2.31) \end{aligned}$$

where (2.28) regards internal links, (2.29) corresponds to internal non-links, (2.30) is external links, and (2.31) corresponds to external non-links. The α_{ij}, β_{ij}, γ_{ij}, η_{ij} are weights of the individual contributions. It is natural to assign the same weight to all links—that is, to choose $\alpha_{ij} = \gamma_{ij}$—and similarly for non-links, $\beta_{ij} = \eta_{ij}$. A convenient choice that allow us to adjust the relative contribution of links and non-links is $\alpha_{ij} = 1 - \theta P_{ij}$ and $\beta_{ij} = \theta P_{ij}$. Here θ is a constant between 0 and 1, and P_{ij} is the probability that a link exists between node i and j normalized such that $\sum_{ij} P_{ij} = 2m$. If $\theta = 1$, this normalization leads to the situation that the weight contributed by links and non-links is equal. This choice of weights simplifies H. We find

$$H = -\sum_{ij}(A_{ij} - \theta P_{ij})\delta(c_i, c_j). \tag{2.32}$$

This equation should remind the reader of the modularity. If we choose[6]

$$P_{ij} = \frac{k_i k_j}{2m}, \tag{2.33}$$

and $\theta = 1$, we find that

$$Q = -\frac{1}{2m}H, \tag{2.34}$$

which makes the connection to Q explicit.

The general framework is only half of Reichardt and Bornholt's achievement. The same authors also point out that the problem of minimizing H is equivalent to determining the ground state of a C-state Potts model. The couplings are $J_{ij} = A_{ij} - \theta P_{ij}$ and exist between all pairs of nodes. The coupling is ferromagnetic where links between nodes exist and it is anti-ferromagnetic, where the links are absent. The letter H is, of course, chosen because the task of finding community structure is equivalent to minimizing a **H**amiltonian of the form equation (2.32).

The observation that community detection is equivalent to finding ground states in spin glass, opens up for the application of many powerful methods that are well known from statistical physics. Again, I will not go into the details of optimization here, since this subject is covered in chapter 6.

[6]Other choices of P_{ij} correspond to different null models. For example, $P_{ij} = p$ would be a good approximation for a random network.

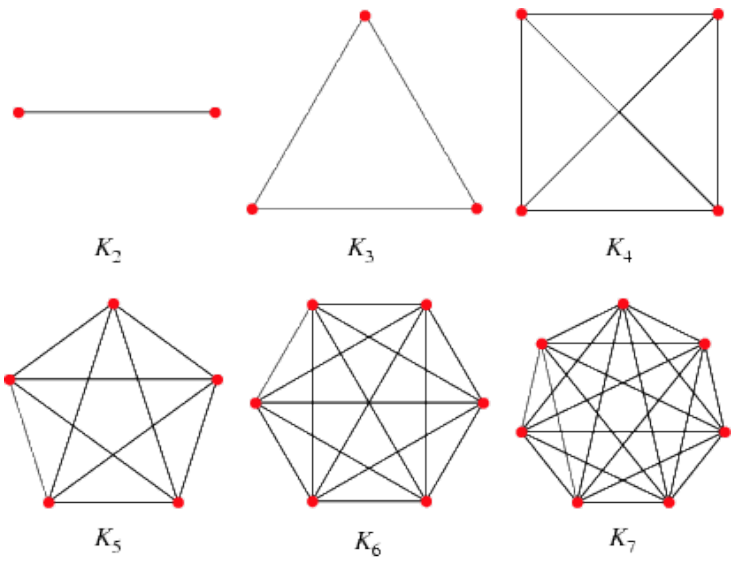


Figure 2.5: Complete graphs. If all pairs of nodes of a certain graph are connected, the graph is said to be *complete*. If a subgraph of a given network is complete, this subgraph is called a *clique*. The size of the clique is identical to the size of the subgraph. This image displays complete graphs of $k = 2,3,4,5,6,7$ nodes. Image from [96]

2.4 k-Cliques

An interesting counterweight to the global optimization algorithms that we have seen in this chapter, is the k-clique community detection algorithm, which is a bottom up method for identifying communities in complex networks [76]. The k-clique algorithm utilizes a certain type of motifs—called *cliques*—to detect communities in unweighted and undirected complex networks. A k-clique is simply a complete subgraph consisting of k nodes, see figure 2.5 for more details.

A community of k-cliques is defined as the union of all k-cliques that can be reached from each other through a series of *adjacent* k-cliques; in this context adjacency is defined as sharing $k-1$ nodes. Figure 2.6 shows a simple example of k-clique communities. One exciting feature of this algorithm, which is evident from figure 2.6, is that it allows for overlap of communities[7]. The k-clique definition of 'community' allows any node to be a member of any number of communities. We know from experience that in a social network, each individual is a member of many different communities: Family, friends, work, hobbies,

[7] Technically, the algorithm by Reichardt and Bornholdt [83] also allows for overlap. This is defined as the case where the ground state is degenerate, that is, if different assignments of nodes to communities lead to the same ground state energy.

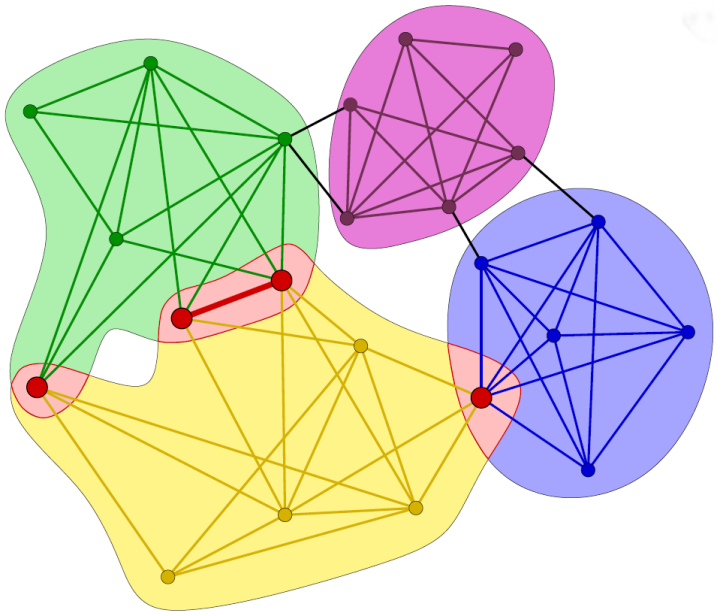


Figure 2.6: An example of cliques, overlap and clique-adjacency. The figure displays four k-cliques with $k = 4$. The k-clique definition allows for overlap between communities. The green and yellow community share 3 nodes, the blue and yellow community share one node; these nodes are said to be *overlapping*. From [76].

etc. Assigning a person to just one of these communities is too simplistic—the concept of overlapping communities is not limited to social networks, but extends to most real world networks. including biological networks, technological, and networks of concepts (see figure 2.7).

Note that once the full set of cliques is located, we must determine an optimal value of k. Tuning this parameter enables a better understanding of the behavior of the network in different levels of detail. If the interest lies in the network around a given node, one would scan (from below) through a range of k's and observe how its communities change—this behavior is highly network-dependent. For high values of k, the network breaks into pieces and only the most-connected cliques remain.

Weighted networks can be analyzed by picking some threshold w^* and keeping all links with weight higher than w^* while discarding all links with a lower weight. Changing this threshold is like changing the resolution (as in a microscope) used to investigate the community structure: In analogy with the variation of k, increasing w^* will cause communities to shrink and fall apart. When observing the network on a global scale, Palla *et al.* [76] suggest a way of setting a global w^*, based on finding a community structure that is as highly structured

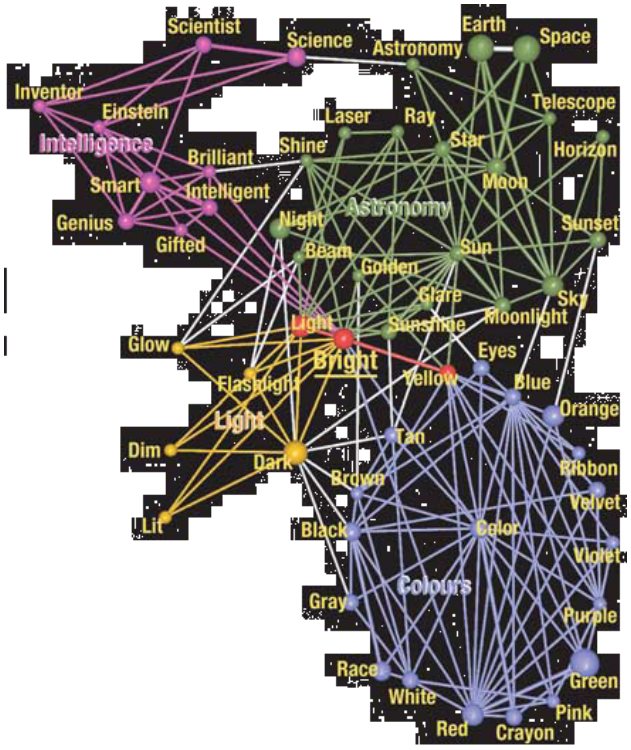


Figure 2.7: Overlapping communities. The communities of the word 'bright' in the South Florida Free Association norms list. Each community is color coded, and the overlapping nodes and links between them are emphasized in red. Finally, the volume of the balls and the width of the links are proportional to the total number of communities that they belong to. Each community represents a different meaning of 'bright'. The network of word meanings is a great example of how every node/word is part of many different communities of meaning. Assigning the word 'bright' to any single category would indicate a fundamental misunderstanding of the nature of the network of words. From [76].

as possible. In the related percolation phenomena, a giant component appears when the number of links is increased above the critical point. Therefore, to approach this critical point from below the threshold w^* is lowered until the largest community is twice as big as the second largest one. This is done for each value of k. The procedure ensures that as many communities as possible are located, without the negative effect of a giant component/community smearing out the details of the community structure by merging many smaller communities. For random networks, the clique percolation behavior is well understood [22, 77].

From these considerations, however, one weakness of the theory becomes apparent. Regardless of their structural relevance, nodes with fewer links than k are never accounted for, because they cannot participate in k-cliques. This also has the consequence that all nodes on the 'periphery' of the network are not part of the analysis. In a network with power-law degree distribution, the majority of nodes often have fewer than $3-4$ links. The opposite problem occurs for the highly connected hubs that tend to obscure the analysis by being

connected to too many other nodes. In this sense, the algorithm works best on networks with normal-type degree distributions and the results are questionable for graphs with power-law degree distribution.

Another potential problem with this algorithm is that the determination of the full set of cliques of a graph is widely believed to be an NP complex problem. Palla *et al.* [76] report, however, that their algorithm is very efficient when applied to the graphs of the investigated real systems. The required CPU time depends very strongly on the structure of the input data. For this reason no closed formula can, in general, be given even to estimate the system size dependence. As an example, it was reported that a complete analysis of a co-authorship network with 127 000 links takes less than 2 hours on a desktop PC anno 2005.

2.5 Status

Although the study of communities has evolved tremendously since Fiedler first developed spectral bisection in 1973 [26], we are still a long way from a coherent and complete description. This state of affairs is emphasized by the fact that no commonly accepted definition of what we mean by a community in a complex network exists.

One obvious contender for a definition of community structure is 'the division of nodes into communities that maximizes the modularity'. This solution, however, suffers from several serious problems. First, it turns out that the factor of $P_{ij} \sim (k_i k_j)/(2m)$ in equation (2.27) has unwanted and unexpected consequences for the division into communities: In a large network, the expected number of links between two small modules is small; therefore, a single link between two such communities is enough to join the two distinct modules into a single community [28]. The normalization by the number of links, m, has the related consequence that if one uses Q-optimization to divide a network into C communities and subsequently adds a disjoint set of nodes to the network and divide the resulting network into $C+1$ communities by optimizing Q, the boundaries between the original C communities can shift substantially compared to the initial division [84]. Finally, as we have discussed above, the current definition of Q suffers from the serious draw-back that it does not allow nodes to belong to more than one community.

The k-clique definition of community structure *does* allow nodes to participate in more than one community. Further, the k-clique definition commences to bridge the gap between the global measures (motifs in section 1.6) and community detection. These attractive traits, however, do not change the fact that the k-clique definition also suffers from many other problems as discussed in section 2.4.

This is the status: Without even a commonly accepted definition[8] of community structure, we are a long way from a satisfying mathematical framework for description of the structure of complex networks. We need a description that is able to bridge the conceptual gap between the tools of chapter 1 and 2, respectively. We also need a definition of community structure that allows for overlap between communities and varying topological hierarchies within each community. In the final chapter of this dissertation, chapter 7, I will present my ideas for the road ahead.

[8]This topic is discussed in the literature. See [81].

CHAPTER 3

SPIRES

THE essential component of any investigation of network structure is, of course, real network data. So far, we have considered the mathematical tools for analyzing network structure. But, without access to data, the tools are meaningless. This chapter concerns the data set, used for most of the work in this dissertation, namely the SPIRES database of scientific papers in high energy physics.

3.1 History of Spires

The SPIRES high energy physics (hep) data base is the oldest computerized data base on the planet. It was founded in 1962, by the Stanford Linear Accelerator Center (SLAC) library and it has been comprehensively collecting hep preprints since then. That same year, Deutsches Elektronen–Synchrotron (DESY) in Hamburg, Germany, began publishing a record called 'High Energy Physics—An Index' (HEPI).

In 1967, computer scientists at Stanford University began working on a new

computerized database that was designed to handle (in principle) a limitless number of large records. Come March 1968, the SLAC Library, being in possession of a large database that was perfect for testing the new system, began participating in this project. Thus, SPIRES (Stanford Physics Information REtrieval System[1]) was born. Due to this origin, SPIRES is also a unique programming language.

In the present context, the fact that the SPIRES database allowed the SLAC librarians to add the reference list of all papers to the data base, is of paramount importance, making the extraction of citation data possible. It was only natural for the DESY and SLAC libraries to cooperate, and by June 1969, the conversion of the DESY data to SPIRES format was complete. By 1974, SLAC and DESY[2] were comprehensively collecting preprints (and by extension published articles) and cataloguing them in a single SPIRES hep database.

The next important step for the SPIRES database was the 1991 creation of the LANL (Los Alamos National Laboratory) e-Print server (now arXiv). This archive allows authors to self-publish their preprints on one common server, assigning, to each paper, a unique number of the form *archive*/0701311. The *archive* label denotes the subfield: For example, astrophysics is *astro-ph*, condensed matter is *cond-mat*, etc., and the number identifies the 311th paper of January 2007. This unique labelling allowed systematic referencing to unpublished papers and now allows citations of preprints to be registered in the SPIRES hep database. This review of the history of SPIRES, is based on a paper by Heath O'Connell [75].

3.2 Information Networks

The network constituted of papers in the SPIRES database is a classical example of an *information network*. The citation network is directed, the nodes are scientific publications and links arise when one papers cites another paper. Incoming links are *citations* and outgoing links are called *references*. To get an intuitive feel for the citation network, recall that the network of papers shares certain prop-

[1]SPIRES was later renamed Stanford Public Information REtrieval System.

[2]Later, CERN, the University of Durham, KEK, the Yukawa Institute, and Fermilab joined the collection of papers.

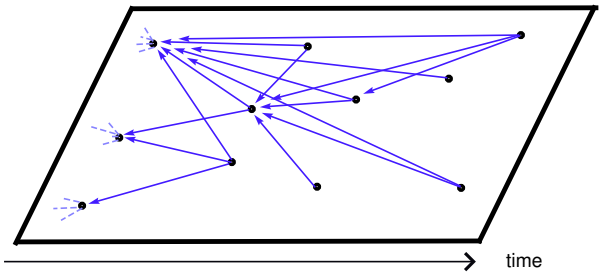


Figure 3.1: An excerpt of the network of papers in SPIRES. The •'s are papers and (directed) links are represented using arrows. Note the time-line. In this illustration, the tree structure is clear; papers can only link back in time, and once they are published, no new outbound links can arise. From [51].

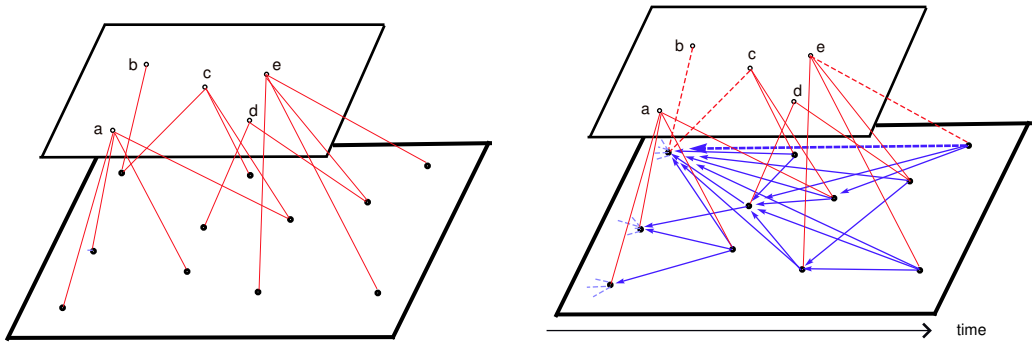
erties with the world wide web. The analogy is: paper ∼ webpage, reference ∼ outgoing link, and citation ∼ incoming link. The structure of the citation network reflects the structure of the information stored at each node, making the network an 'information network'. Scientific papers are printed on paper, and therefore the citation network instantly freezes into a tree structure, when a paper is published, cf. figure 3.1. The network of scientific papers contains no cycles.

Another important example of an information network is the world wide web. Unlike the citation network, the world wide web is cyclic, simply because there is no natural ordering of sites and no constraints to prevent the appearance of closed loops. The world wide web has been under continuous scrutiny since its appearance coincided with the emergence of complex network science in the early 1990s. The studies by Albert *et al.* [6, 11], Kleinberg *et al.* [44], and Broder *et al.* [17], were particularly influential.

Other examples of information networks are the network of citation between U.S. patents [38] and peer-to-peer (P2P) networks, which are networks between computers that allow file sharing [2, 3]. Finally, the network between word classes in a thesaurus can be thought of as an information network (where one 'surfs' from meaning to meaning) [42, 45]—although this can also be regarded as a conceptual network representing the structure of the language [63] (cf. figure 2.7).

3.3 Longitudinal Correlations

Including knowledge about the author of every publication adds a new level of complexity to the network we have been considering the previous section. So far, we have neglected the structure that emerges when we include this new level in the investigation of the citation network. From now on, let us use the distinction that the *author network* refers to the entire network, including the author information, whereas the *paper network* refers to the network of papers discussed in the previous chapter. A clear understanding of the details of the author network is of great importance, when we include the author correlations in the analysis in chapter 5.



(a) (b)

Figure 3.2: A visualization of the author network. (a) Displays (a small portion of) the author level connecting to (an even smaller portion of) the paper level; each author is represented by a '∘' and a letter from a to e, each paper is represented by a '•'. (b) Here, the paper-network from figure 3.1 has been added to the picture. This representation provides an excellent illustration of how the author-level induces correlations on the level of papers; see the main text for further details. From [51].

The most natural way to represent the author network visually, is to use two levels, consider figure 3.2. The (upper) level of authors in figure 3.2 (a) connects to the (lower) paper level by means of their publications. Each of the 5 authors (a – e) have authored a number of publications represented by '•'s on the paper level. In figure 3.2 (b) the directed network of citations and references between papers that was discussed in the previous chapter, has been added to the lower level, cf. figure 3.1.

Analogously to the definition of citations and references from the previous chap-

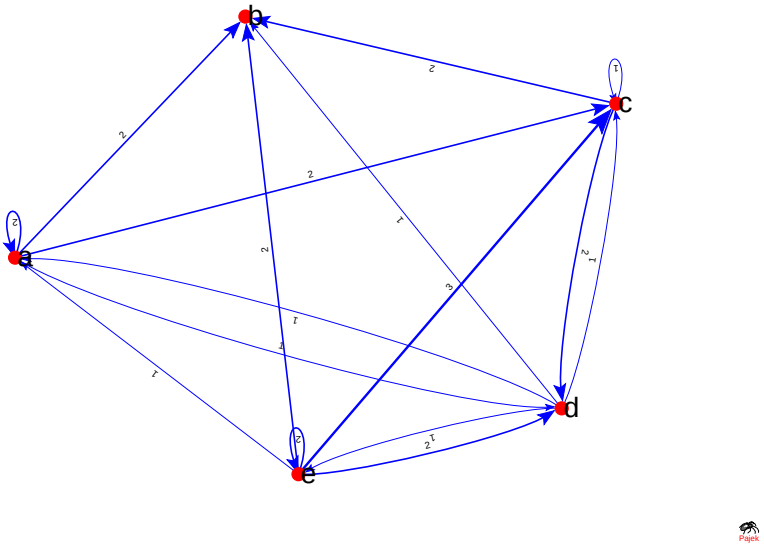


Figure 3.3: A visualization of the author network from figure 3.2 (b) collapsed into one level of authors citing authors; each line is labelled by the number of citations it represents, also the thickness of a line is proportional to the number of citations it represents. Loops signify self citations. This network was generated using Pajek network visualization software [92]. From [51]

ter, we define an *author*'s references as the papers listed at the end of *all* of his papers and correspondingly, his citation count is the cumulated sum of citations in this entire citation record. This definition and the two level representation underscores that *references and citations between authors run via the paper network*. As an example of this, consider figure 3.2 (b), where the dashed line illustrates how author e cites authors b and c via a reference from one of e's papers to a (highly cited) paper co-authored by b and c. Sometimes, it is convenient to disregard the fact that citations between authors run via the network of papers, and collapse the two levels into one single level of authors citing other authors.

For example, such a collapse of the network in panel (b) of figure 3.2 so that only the links between authors are visible, results in the structure displayed in figure 3.3. This figure is confusing because so much structure is repressed: Here, a number is affiliated with each edge, the graph contains loops, etc.—this representation of the network also makes the time-line from figure 3.2 (b) an impossibility. Generally, the two-level representation strengthens our intuitions about

the structure of the author network.

Finally, the author-network can also be thought of as an unweighted network with links defined by either co-authorship or citations. In such a network, each node can be characterized by a number of properties. Specifically, the network where each node (author) is characterized by a list of papers each written by the author and with with degrees $\{k_i\}$, will be considered in chapter 5.

Previously, we have discussed the similarities between the paper network and the world wide web. The inclusion of the author level in the considerations, however, reveals that the two networks are, in fact, radically different; the internet does not possess any structural property analogous to the strong correlations that the author level imposes on the network of papers. This fact has been almost completely ignored in the literature. Here the citation network is usually considered a much simpler network than the world wide web, due to the tree-structure and acyclicity, cf. the discussion of correlations in the simple growth model in section 1.4.

3.4 Further Reading

For more details on the actual data in SPIRES, the reader is directed to the work by Lehmann [51] and Lehmann *et al.* [58].

Part II

The Papers

CHAPTER 4

Modelling

ONE of the surprising insights gained from the careful study of SPIRES by Lehmann *et al.* [51,58] is the fact that a surprisingly large fraction (29%) of the high energy physics papers are never cited. And 29% is the number obtained without correcting for self-citations; the removal of self-citations would make the fraction of uncited papers substantially higher. At the other end of the food-chain, the 4% most cited authors account for 50% of the citations in the data base. Lehmann *et al.* continue:

> While it is a truism that progress in physics is driven by a few great minds, it can be disturbing to confront this quantitatively. The picture which emerges is thus a small number of interesting and significant papers swimming in a sea of dead papers. This has the practical consequence that any study seeking to understand the dynamics of interesting papers will be forced to discard most papers and accept the greatly increased statistical uncertainties." [58]

The rather striking image of a few living papers swimming in a sea dead papers, is the outset of a rather successful modelling effort [53, 54], described below.

4.1 Life, Death, and Preferential Attachment

The model by Lehmann *et al*. [54] provides a description the citation network on the paper level, cf. section 3.2. The model describes a growing network of citations, similar to the situation presented in section 1.4. However, motivated by a significant inhomogeneity in the data, this model explores the consequences of a distinguishing between 'live' and 'dead' network nodes. 'live' nodes are able to acquire new links whereas 'dead' nodes are static. Lehmann *et al*. develop an analytically soluble augmentation of the growing network model described in section 1.4 that incorporates the distinction between living and dead papers: Each paper (node) enters the data base 'live', but has a probability of dying at each time step that is inversely proportional to its number of citations (links).

It is demonstrated that the live-dead model provides an excellent description of the empirical degree distributions of both live and dead papers in SPIRES. An excellent fit, with remarkable little strain, is obtained for both of these distributions using only three parameters. All of these parameters have immediate physical interpretations: The mean number of citations for live and dead papers, and the fraction of dead papers. All values of these parameters are determined by the fit to the distribution, and the numerical values found are in excellente agreement with the actual data.

Furthermore, Lehmann *et al*. demonstrate that the death mechanism alone can result in power law degree distributions for the resulting network.

Europhys. Lett., **69** (2), pp. 298–303 (2005)
DOI: 10.1209/epl/i2004-10331-2

Life, death and preferential attachment

S. Lehmann[1], A. D. Jackson[2] and B. Lautrup[2]

[1] *Informatics and Mathematical Modeling, Technical University of Denmark Building 321 - DK-2800 Lyngby, Denmark*
[2] *The Niels Bohr Institute - Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark*

received 17 September 2004; accepted 10 November 2004
published online 22 December 2004

PACS. 89.65.-s – Social and economic systems.
PACS. 89.75.-k – Complex systems.

Abstract. – Scientific communities are characterized by strong stratification. The highly skewed frequency distribution of citations of published scientific papers suggests a relatively small number of active, cited papers embedded in a sea of inactive and uncited papers. We propose an analytically soluble model which allows for the death of nodes. This model provides an excellent description of the citation distributions for live and dead papers in the SPIRES database. Further, this model suggests a novel and general mechanism for the generation of power law distributions in networks whenever the fraction of active nodes is small.

That progress in science is driven by a few great contributions becomes disturbingly clear when one considers citation statistics. The vast majority of scientific papers is either completely unnoticed or minimally cited. In high-energy physics, 4% of all papers account for 50% of the citations, while 29% of all papers are not cited at all [1].

In a pioneering sociological work analyzing American high-energy physicists, Cole and Cole [2] connect this high degree of stratification in the scientific literature to what they call *cumulative advantage*. The concept underlying cumulative advantage was originally introduced by Merton [3] with the more striking name of the "*Matthew Effect*". Merton's simple observation was that success seems to breed success. A paper which has been cited many times is more likely to be cited again than one which is less cited, since "unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken away even that which he hath" [4] —hence the name.

Inspired by refs. [2, 3] and his own work on citation networks [5], de Solla Price recast Simon's [6] ideas on the mathematics leading to the power law distributions found in nature and society into the first mathematical model of a scale-free network [7]. Much later, the principles underlying Price's model were independently re-discovered by Barabási and Albert [8], who coined yet another name for the same effect, namely *preferential attachment*, and also firmly established the field of network theory as a branch of physics, cf. reviews in refs. [9–11]. Preferential attachment has since become a widely accepted explanation of the power law degree distributions in complex networks in general. The strength of the preferential attachment model in either incarnation is its simplicity, but this can also be its weakness. In particular, such models tend to assume that networks are homogeneous. When real-world networks can be shown to have identifiable and significant inhomogeneities, preferential attachment must be supplemented by appropriate additional ingredients.

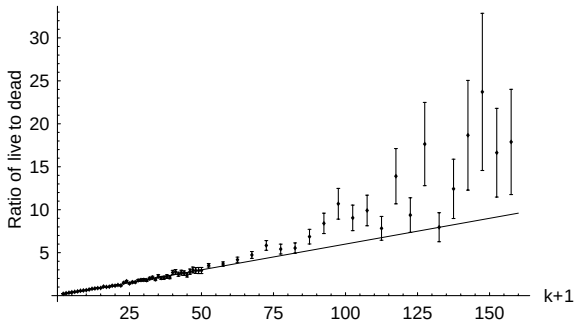


Fig. 1 – The ratio of live to dead papers. The solid straight line has been inserted to illustrate the linear relationship between the live and dead populations for low values of k. The error bars are calculated from the square roots of the citation counts.

For example, it is an empirical fact that the vast majority of nodes in citation networks "die" after a relatively short time and are never cited again. A relatively small population of papers remains alive and continues to accumulate citations many years after publication; this is the main conclusion in ref. [1]. The distinction between live and dead populations represents an important inhomogeneity in the citation data that is not considered in the simple preferential-attachment model. We do not suggest that the presence of death in citation networks diminishes the importance of preferential attachment; however, the distinctly different citation distributions observed for live and dead papers compel us to include the effects of the death of papers in our modeling efforts. It is the purpose of this paper to suggest one such extension of preferential attachment models.

Dead papers. – The work in this paper is based on data obtained from the SPIRES database of papers in high-energy physics. To be specific, the data used below is the network of all citable papers from the Theory subfield of SPIRES, ultimo October 2003. Filtering out all papers for which no information of publication time is available, we are left with a network of 275665 nodes (*i.e.*, papers). All citations to papers not in this network were removed, resulting in 3434175 edges (*i.e.*, citations).

Clearly, there is a variety of ways to define what is meant by a dead node in real data([1]). We have tested several definitions, and our results are qualitatively independent of the specifics of the definition. We have chosen to define papers that have not been cited in 2003 to be dead. Having identified a population of dead papers, we have determined the citation distributions for live and dead papers. These distributions are shown in fig. 2(a) and indicate that the two distributions are significantly different. As suggested in the introduction, most (*i.e.*, approximately three-quarters) of the papers in SPIRES are dead. It is also a simple matter to determine the empirical ratio of live to dead papers as a function of the number of citations per paper k. Figure 1 displays this ratio in the range $1 \leq k \leq 150$. Over most of this range the data is described by a straight line. We note that the data for dead papers with high k-values is very sparse. Since only 0.15% of dead papers have more than 100 citations, statistics beyond this point are highly unreliable. Thus, plotting the ratio of live to dead papers gives a pessimistic representation of the data. The ratio of dead to live papers is described satisfactorily by the simple form $b/(k+1)$ for all but the highest values of k, where this form overestimates the number of dead papers by a factor of two to three. In short, fig. 1

([1])We recognize that there are examples of papers that receive new citations after a long dormant period. However, such cases are rare and do not affect the large-scale statistics.

implies that —to a fairly good approximation— the fraction of dead papers with k citations is proportional to $1/(k+1)$. We will make use of this fact in the next section to suggest an extension of the preferential attachment model which includes the effects of death.

Modeling death and preferential attachment. – Following the usual structure of preferential attachment models, we imagine that at every update a new paper makes m references to papers already in the network and then enters the network with $k = 0$ real citations and $k_0 = 1$ "ghost" citations. Since we have chosen to eliminate all references to papers not in SPIRES in constructing our data set, there is an obvious and rigorous sum rule that the average number of citations per paper is also m. The probability that a paper in the network will receive one of these references is assumed to be proportional to its current total of real and ghost citations. We can estimate when the effects of preferential attachment become important by regarding k_0 as a free parameter. Since we see no *a priori* reason why a paper with 2 citations should have a significant advantage in acquiring citations over a paper with 1 citation, we prefer to allow the data to decide. Thus, in our model, the probability that a paper with k citations acquires a new citation at each time step is proportional to $k + k_0$ with $k_0 > 0$. We can think of the displacement, k_0, as offering a way to interpolate between full preferential attachment ($k_0 = 1$) and no preferential attachment ($k_0 \to \infty$), cf. [12].

More importantly, at every update each live paper in the network has some probability of dying. Guided by the SPIRES data, we assume that this probability is proportional to $1/(k+1)$ for a paper with k real citations. Once dead, a paper can no longer receive new citations. In his 1976 paper, Price notes that cumulative advantage is only half the Matthew Effect, because although success is rewarded, there is no punishment for failure. In this sense, the model described here represents one implementation of the *full* Matthew Effect. Since the rate at which papers are killed is inversely proportional to the number of citations which they have, low cited papers have a much higher probability of paying the ultimate penalty.

The rate equation approach introduced in the context of networks by Krapivsky, Redner, and Leyvraz [13] can easily be modified to allow for death. We let L_k be the probability for finding a live paper with k citations and D_k be the probability of finding a dead paper with k citations. Each paper cites m other papers in the database. Papers are loaded into the database with in-degree $k = 0$. We arrive at the following rate equations:

$$L_k = m(\lambda_{k-1} L_{k-1} - \lambda_k L_k) - \eta_k L_k + \delta_{k,0}\,, \tag{1}$$

$$D_k = \eta_k L_k\,, \tag{2}$$

where λ_k and η_k are rate constants. We define L_k to be equal to zero for $k < 0$ and since every paper has a finite number of citations, the probabilities L_k must become exactly zero for sufficiently large k. Thus, we can let all sums run from $k = 0$ to infinity. While the total citation distribution is, of course, given by $L_k + D_k$, we can also probe the live and dead distributions separately both theoretically and empirically. For any choice of λ_k and η_k, these equations trivially satisfy the normalization condition on the total distribution. However, the constraint that the mean number of references equals the mean number of citations, $\sum_k k(L_k + D_k) = m$, must be imposed by an overall scaling of the λ_k and η_k. Equation (2) shows that the coefficients, η_k, are simply the ratio of dead to live papers as a function of k. Given the empirical values of this ratio shown in fig. 1, our model corresponds to the case where

$$m\lambda_k = a(k + k_0) \qquad \text{and} \qquad \eta_k = \frac{b}{k+1}\,. \tag{3}$$

Performing the recursion, we find

$$L_k = \frac{\Gamma(k+2)}{ak_1k_2}\frac{\Gamma(k+k_0)}{\Gamma(k_0)}\frac{\Gamma(1-k_1)}{\Gamma(k-k_1+1)}\frac{\Gamma(1-k_2)}{\Gamma(k-k_2+1)}\,, \tag{4}$$

where k_1 and k_2 are the solutions to the quadratic equation

$$(a(k+k_0)+1)(k+1)+b=0 \tag{5}$$

regarded as a function of k.

One general observation of some interest emerges in the limit $k_0 \to \infty$ in which preferential attachment is turned off. We obtain this limit by making the replacement $\alpha = ak_0$ in eq. (4) and then taking the limit $k_0 \to \infty$ for fixed α. A little work reveals that

$$L_k = \frac{1}{\alpha}\left(\frac{\alpha}{1+\alpha}\right)^{k+1}\frac{(\frac{b}{1+\alpha})!(k+1)!}{(\frac{b}{1+\alpha}+k+1)!}\,. \tag{6}$$

The D_k are simply $bL_k/(k+1)$ as before. (Equation (6) can also be obtained by solving eqs. (1) and (2) with constant λ_k and $\eta_k = b/(k+1)$; the two approaches are equivalent.) When the death mechanism is eliminated by setting $b=0$, the resulting distribution shows an exponential decrease which is to be expected given the assumed absence of preferential attachment.

In fact, the death of nodes offers an alternative mechanism for obtaining power laws. To see this, consider the limit $\alpha \to \infty$ and $b \to \infty$ with the ratio $r = b/(\alpha+1) \approx b/\alpha$ fixed. In this limit it is tempting to replace the term $\alpha/(1+\alpha)$ by 1, which allows us to compute simple expressions for the fraction of dead papers f and the average number of citations of the live and dead papers, $m_{\rm L}$ and $m_{\rm D}$. (This approximation is appropriate when $r \geq 2$. When $r < 2$ the neglected factor is essential for ensuring the convergence of $m_{\rm L}$ and/or $m_{\rm D}$.) The fraction of live papers is then

$$1 - f = \frac{1}{\alpha(r-1)}\,, \tag{7}$$

and the average number of citations for the live papers and dead papers, respectively, is

$$m_{\rm L} = \frac{2}{r-2} \qquad \text{and} \qquad m_{\rm D} = \frac{1}{r-1}\,. \tag{8}$$

The average number of citations for all papers is evidently $m_{\rm D}$ in the limit $\alpha \to \infty$ for which $f \to 1$. Most importantly, we see in this limit that

$$L_k \sim \frac{1}{k^r} \qquad \text{and} \qquad D_k \sim \frac{b}{k^{r+1}} \tag{9}$$

for $k > r$. Thus, we see that power law distributions for both live and dead papers emerge naturally in the limit where the fraction of dead papers f goes to 1. In this limit, a vanishing fraction of live papers swim in a sea of dead papers. Since such power laws are sometimes regarded as an indication of preferential attachment, it is useful to see a quite different way of obtaining them.

Death in the real world. – We now return to the full model and compare it to the data from SPIRES. If we assign all zero cited papers to the dead category, the mean number of citations is 34.1 for live papers, 4.5 for dead papers, and 12.5 for all papers. The fraction of live papers is 27.0%. By minimizing the squared fractional error, we can fit the live data with an rms error of only 21% using the forms of eqs. (4) and (5) with the parameters $k_0 = 65.6$,

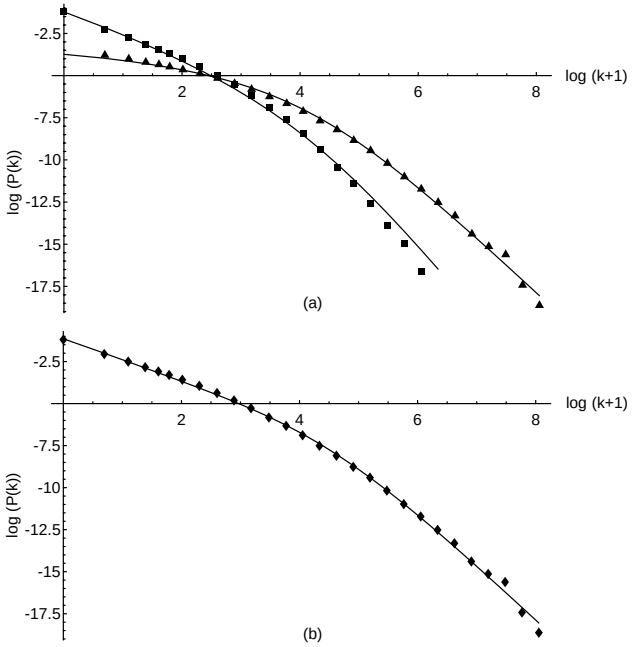


Fig. 2 – (a) Log-log plots of the distributions for live and dead papers. The triangles are the live data and the squares are the dead data. The solid lines are the fit. (b) A log-log plot of the distribution of all papers (live plus dead). The points are the data; the solid lines are the fit.

$a = 0.436$, and $b = 12.4$. Given that the data spans six orders of magnitude, the quality of this agreement is strikingly high. The results of the fits are displayed in fig. 2.

The fitted mean number of citations is 32.9 citations for live papers, 4.25 for dead papers, and 12.8 for all papers. According to the fit, 7.5% of all papers with 0 citations are, in fact, alive. Assigning this fraction of zero citation papers to the live data, we find mean citations of 31.5, 4.6, and 12.5 respectively. We also find that 29.2% of the papers in the model are live. This is in excellent agreement with the data. There is remarkably little strain in the fit. We can, for example, determine the model parameters a, b, and k_0 from the empirical values of m_{L}, m_{D}, and f. This leads to small changes in the model parameters and yields a description of comparable quality for the distributions. It is clear from fig. 2 that the present fit to the live distribution leads to some systematic errors in the description of the dead population for the highest values of k. Given the deviations from a straight line of the data of fig. 1 for large k, this comes as no surprise. This could obviously be remedied by a small modification of the η_k through the inclusion of a suitable k^2-term in the denominator.

It is clear that the present simple model is capable of fitting the distributions of both live and dead papers with remarkable accuracy. We note that the best-fit value of the parameter $k_0 = 65.6$ suggests that a paper with $k = 66$ citations has a competitive advantage over a paper with no citations of a factor of 2 rather than the factor of 67 suggested by the simplest preferential attachment models.

Discussion and conclusions. – It is obvious that the death mechanism introduced here is essential if we wish to consider the empirical citation distributions of live and dead papers separately. It is less obvious that the death mechanism (*i.e.*, $b \neq 0$) is required to provide a good description of the total citation data. A similar fit to the citation distribution for all papers

with the constraint $b = 0$ yields the parameters $a = 0.528$ and $k_0 = 13.22$ and gives an rms fractional error of 33.6%. Although there are some indications of systematic deviations in the resulting fit, its overall quality remains high in spite of the fact that this constrained fit ignores important correlations present in the data set. This result illustrates the familiar fact that more detailed modeling is not necessarily required to fit global network distributions even if important empirical correlations are neglected in the process. It also reminds us of the equally familiar corollary that even a high-quality fit to global network distributions cannot safely be regarded as an indication of the absence of additional correlations in the data. The most significant difference between the model parameters obtained with and without the death mechanism is the value of k_0, which changes by a factor of 5 from 65.6 to 13.2. We have an intuitive preference for the larger value. (We believe that preferential attachment will play an important role when a paper is sufficiently visible that authors feel entitled to cite it without reading it and that $k_0 \approx 65$ represents a reasonable threshold of visibility.) It is clear, independent of such subjective preferences, that it is dangerous to assign physical significance to even the most physically motivated parameters if a network contains unidentified correlations or if known correlations are neglected in the modeling process. Specifically, it is difficult to draw firm conclusions regarding the onset of preferential attachment if the death mechanism is not included.

We have identified significant differences between the citation distributions of live and dead papers in the SPIRES data, and we have constructed a model including both modified preferential attachment and the death of nodes that is quantitatively successful in describing these differences. We have further seen that the death mechanism can provide an alternate mechanism for producing power law distributions when the fraction of live nodes is small. Since many networks involve a small fraction of active nodes, this mechanism may be of more general utility. However, the numerical success of the present model does not indicate the absence of additional correlations in the SPIRES data. In fact, we know that such correlations exist. Consider the conditional probability, $P(k|\bar{m})$, that a paper written by an author with a lifetime average of $\bar{m}$ citations per paper will receive k citations. The general interest in citation data is based on the widespread intuitive belief that $P(k|\bar{m})$ is a sensitive function of $\bar{m}$. This belief is supported by the SPIRES data and will be treated in a subsequent publication.

* * *

Our grateful thanks to T. C. Brooks at SPIRES without whose swift replies and thoughtful help we would have lacked all of the data!

REFERENCES

[1] Lehmann S., Jackson A. D. and Lautrup B. E., *Phys. Rev. E*, **68** (2003) 026113.
[2] Cole J. R. and Cole S., *Social Stratification in Science* (The University of Chicago Press, London, Chicago) 1973.
[3] Merton R. K., *Science*, **159** (1968) 56.
[4] Matthew, xxv, 29 (The Holy Bible, King James version).
[5] Derek de Solla Price, *Science*, **149** (1965) 510.
[6] Simon H. A., *Models of Man* (Wiley, New York) 1957.
[7] de Solla Price Derek, *J. Am. Soc. Information Sci.*, **27** (1976) 292.
[8] Barabási A.-L. and Albert R., *Science*, **286** (1999) 509.
[9] Newman M. E. J., *SIAM Rev.*, **45** (2003) 167.
[10] Dorogovtsev S. N. and Mendes J. F. F., *Adv. Phys.*, **51** (2002) 1079.
[11] Albert R. and Barabási A.-L., *Rev. Mod. Phys.*, **74** (2002) 47.
[12] Krapivsky P. L. and Redner S., *Phys. Rev. E*, **63** (2001) 066123.
[13] Krapivsky P. L., Redner S. and Leyvraz F., *Phys. Rev. Lett.*, **85** (2000) 4629.

Cited by

The paper *Life, Death, and Preferential Attachment* has been cited by the following papers [15, 36, 46, 48, 53, 55, 82]

4.2 Live and Dead Nodes

Live and Dead Nodes [53] originated as a contribution to *SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security (April 2005)*, but the paper was later published [53]. It covers much of the same ground as *Life, Death, and Preferential Attachment* from section 4.1, but it contains some additional analytical considerations, regarding the form of the degree distribution.


Computational & Mathematical Organization Theory, 11, 161–170, 2005

Live and Dead Nodes

S. LEHMANN
Technical University of Denmark, Informatics and Mathematical Modeling, Building 321, DK-2800 Kgs, Lyngby, Denmark
email: lehmann@nbi.dk

A.D. JACKSON
The Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark
email: jackson@nbi.dk

Abstract

In this paper, we explore the consequences of a distinction between 'live' and 'dead' network nodes; 'live' nodes are able to acquire new links whereas 'dead' nodes are static. We develop an analytically soluble growing network model incorporating this distinction and show that it can provide a quantitative description of the empirical network composed of citations and references (in- and out-links) between papers (nodes) in the SPIRES database of scientific papers in high energy physics. We also demonstrate that the death mechanism alone can result in power law degree distributions for the resulting network.

Keywords: power law, network evolution, networks, network models, citations

1. Introduction

The study and modeling of complex networks has expanded rapidly in the new millennium and is now firmly established as a science in its own right (Watts, 1999; Albert and Barabási, 2002; Dorogovtsev and Mendes, 2002; Newman, 2003). One of the oldest examples of a large complex network is the network of citations and references (in- and out-links) between scientific papers (nodes) (de Solla Price, 1965; Redner, 1998; Lehmann et al., 2003, 2005; Redner, 2004). A very successful model describing networks with power-law degree distributions is based on the notion of *preferential attachment*. The principles underlying this model were first introduced by Simon (Simon, 1957), applied to citation networks by de Solla Price (de Solla Price, 1976),[1] and independently rediscovered by Barabási and Albert (Barabási and Albert, 1999). Various modifications of the preferential attachment model have appeared more recently. In the present context, the key papers on preferential attachment are Lehmann et al. (2003, 2005), Krapivsky et al. (2000), Krapivsky and Redner (2001) and Klemm and Eguíluz (2002). Simplicity is both the primary strength and the primary weakness of the preferential attachment model. For example, preferential attachment models tend to assume that networks are homogeneous. When networks have significant and identifiable inhomogeneities (as is the case for the citation network), the data can require augmentation of the preferential attachment model to account for them.

The primary conclusion of Ref. (Lehmann et al., 2003) is that the majority of nodes in a citation network 'die' after a short time, never to be cited again. A small population of papers remains 'alive' and continues to be cited many years after publication. In Ref. (Lehmann et al., 2005) it was established that this distinction between live and dead papers is an important inhomogeneity in the citation network that is not accounted for by the simple preferential attachment model. Interestingly, a similar distinction between live and dead nodes was recently independently suggested by Redner (2004). In this paper, we will explore how the distinction between live and dead papers manifests itself in network models and thus suggest an extension of the preferential attachment model.

2. The SPIRES Data

The work in this paper is based on data obtained from the SPIRES[2] database of papers in high energy physics. More specifically, our dataset is the network of all citable papers from the theory subfield, ultimo October 2003. After filtering out all papers for which no information of time of publication is available and removing all references to papers not in SPIRES, a final network of 275 665 nodes and 3 434 175 edges remains.

Above we described a dead node as one that no longer receives citations, but how does one define a dead node in *real* data? We have tested several definitions, and the results are qualitatively independent of the definition chosen. Therefore, we can simply define live papers as papers cited in 2003. While we acknowledge the existence of papers that receive citations after a long dormant period, such cases are rare and do not affect the large scale statistics. In figure 2, the (normalized) degree distributions of live and dead papers in the SPIRES data are plotted, and it is clear that the two distributions differ significantly. Having isolated the dead papers, we are not only able to plot them; we can also determine the empirical ratio of live to dead papers as a function of the number of citations per paper, k. In figure 1 this ratio is displayed with k ranging from 1 to 150 (papers with zero citations are dead by definition). Over most of this range, the data is well described by a straight line. Note that the data for dead papers with high citation counts is very sparse. For example, only 0.15% of the dead papers have more than 100 citations, so the statistics beyond this point are highly unreliable. More generally, a linear plot of the ratio of live to dead papers provides a pessimistic representation of the data. We therefore conclude that the ratio of *dead to live* papers is relatively well described by the simple form $1/(k+1)$ for all but the largest values of k, for which the number of dead papers is overestimated by a factor of two to three. In the following section, we will make use of this relation to extend the preferential attachment model to include dead nodes.

3. The Model

The basic elements of the preferential attachment model are *growth* and *preferential attachment* (Barabási and Albert, 1999). The simplest model starts out with a number of initial nodes and at each update, a new node is added to the database. Each new node has m outlinks that connect to the nodes already in the database. Each new node enters with $k = 0$

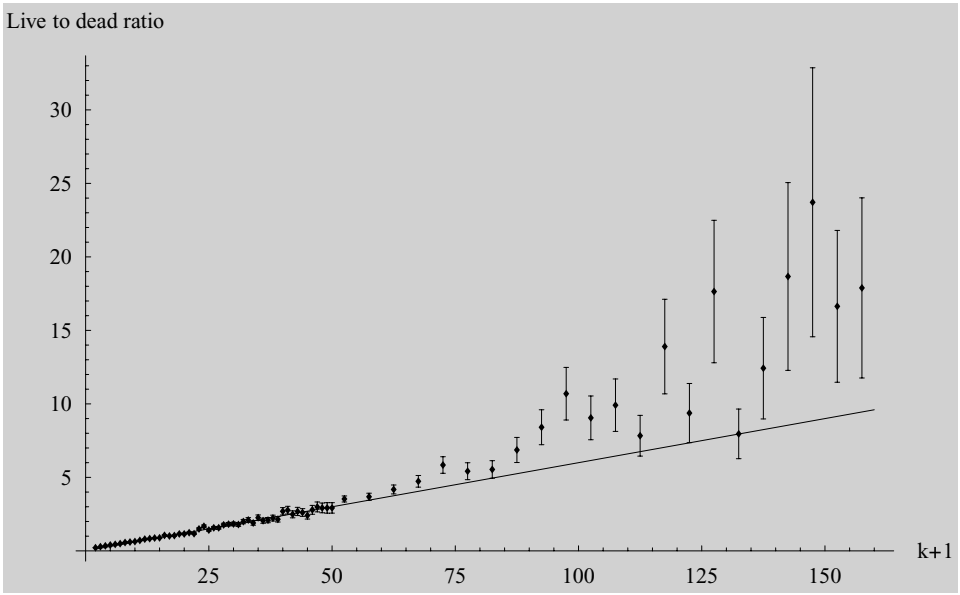


Figure 1. Displayed above is ratio of live to dead papers as a function of k. Error bars are calculated from square roots of the citation counts in each bin. Also, a straight line is present to illustrate the linear relationship between the live and dead populations for low values of k.

real in-links. This is the *growth* element of the model. Note that, since we have chosen to eliminate all references to papers not in SPIRES from the dataset, there is a sum rule such that the average number of citations per paper is also m. *Preferential attachment* enters the model through the assumption that the probability for a given node already in the database to receive one of the m new in-links is proportional to its current number of in-links. In order for the newest nodes (with $k = 0$ in-links) to be able to begin attracting new citations, we load each node into the database with $k_0 = 1$ 'ghost' in-links that can be subtracted after running the model. The probability of acquiring new citations is proportional to the *total* number of in-links, both real and ghost in-links.

One of the simplest ways to augment this simple incarnation of the preferential attachment model described above is to regard k_0 as a free parameter. This allows us to estimate when the effects of preferential attachment become important. Since there is no a priori reason why a paper with 2 citations (in-links) should have a significant advantage over a paper with 1 citation, it is preferable to let the data decide. Thus, in our model, the probability that a live paper with k citations acquires a new citation at each time step is proportional to $k + k_0$ with $k_0 > 0$. Also, note that we can think of the displacement k_0 as a way to interpolate between full preferential attachment ($k_0 = 1$) and no preferential attachment ($k_0 \to \infty$).

The significant extension of the simple model to be considered here is that, in our model, *each paper has some probability of dying at every time step*. From Section 2, we have a

very good idea of what this probability should be: Figure 1 shows us that for a paper with k citations, this probability is proportional to $1/(k+1)$ to a reasonable approximation. With this qualitative description of the model in hand, we proceed to its solution.

4. Rate Equations

One very powerful method for solving preferential attachment network models is the rate equation approach, introduced in the context of networks by Krapivsky et al. (2000). Let L_k and D_k be the respective probabilities of finding a live or a dead paper with k real citations. As explained above, we load each paper into the database with $k = 0$ real citations and m references. The rate equations become

$$L_k = m(\lambda_{k-1}L_{k-1} - \lambda_k L_k) - \eta_k L_k + \delta_{k,0} \tag{1}$$

$$D_k = \eta_k L_k, \tag{2}$$

where λ_k and η_k are rate constants. Since every paper has a finite number of citations, the probabilities L_k and D_k become exactly zero for sufficiently large k; we also define L_k to be zero for $k < 0$. In this way, all sums can run from $k = 0$ to infinity. These equations trivially satisfy the normalization condition

$$\sum_k (L_k + D_k) = 1, \tag{3}$$

for any choice of η_k and λ_k. However, we also demand that the mean number of references is equal to the mean number of papers

$$\sum_k k(L_k + D_k) = m. \tag{4}$$

This constraint must be imposed by an overall scaling of η_k and λ_k. The model described in Section 3 corresponds to a choice of η_k and λ_k, where

$$m\lambda_k = a(k + k_0) \tag{5}$$

is the preferential attachment term and

$$\eta_k = \frac{b}{k+1} \tag{6}$$

corresponds to the previously described death mechanism. We insert Eqs. (5) and (6) into Eq. (1) and perform the recursion to find

$$L_k = \frac{\Gamma(k+2)}{ak_1k_2}\frac{\Gamma(k+k_0)}{\Gamma(k_0)}\frac{\Gamma(1-k_1)}{\Gamma(k-k_1+1)}\frac{\Gamma(1-k_2)}{\Gamma(k-k_2+1)}, \tag{7}$$

and of course $D_k = bL_k/(k+1)$. The two new constants, k_1 and k_2 are solutions to the quadratic equation

$$(a(k+k_0)+1)(k+1)+b=0 \tag{8}$$

as a function of k.

5. The $k_0 \to \infty$ Limit

Before moving on, let us explore the limit where $k_0 \to \infty$ and preferential attachment is turned off. In this regime, the network is, of course, completely dominated by the death mechanism. We can either obtain this limit by again solving Eqs. (1) and (2) with $\lambda_k =$ *constant* and $\eta_k = b/(k+1)$, or we can make the more elegant replacement $\alpha = ak_0$ in Eq. (7), and then take the limit $k_0 \to \infty$ for fixed α. The two approaches are equivalent. We find

$$L_k = \frac{1}{\alpha}\left(\frac{\alpha}{1+\alpha}\right)^{k+1} \frac{\left(\frac{b}{1+\alpha}\right)!(k+1)!}{\left(\frac{b}{1+\alpha}+k+1\right)!}, \tag{9}$$

and the D_k are still simply $bL_k/(k+1)$. With this expression for L_k, let us consider the limit of $\alpha \to \infty$ and $b \to \infty$ with the ratio $r = b/(\alpha+1) \approx b/\alpha$ fixed. In this limit, it is tempting to replace the term $\alpha/(\alpha+1)$ by one.[3] In this case, the use of identities, such as

$$\sum_{k=1}^{\infty} \frac{k!}{(k+r)!} = \frac{1}{(1-r)r!} \tag{10}$$

enable us to compute the fraction of dead papers f, and the average numbers of citations for live and dead papers. The results are simply

$$1-f = \frac{1}{\alpha-1} \tag{11}$$

$$m_L = \frac{2}{r-2} \tag{12}$$

$$m_D = \frac{1}{r-1}, \tag{13}$$

and the average number of citations for all papers is evidently $m = (1-f)m_L + fm_D$. The fraction of dead papers is $f \to 1 - \mathcal{O}(1/b)$ and the average number of citations for all papers approaches m_D.

The most important result, however, is that in this limit we find that

$$L_k \sim \frac{1}{k^r} \quad \text{and} \quad D_k \sim \frac{b}{k^{r+1}}, \tag{14}$$

where we assume that $k > r$. Thus, *we see that power law distributions for both live and dead papers emerge naturally in the limit of* $f \to 1$. In the literature, power laws in the degree distributions of networks are often regarded as an indication that preferential attachment has played an essential part in the generation of the network in question. It is thus of considerable interest to see an alternative and quite different way of obtaining them.

6. The Full Model

Let us now return to the full model and see how it compares to the data from SPIRES. With all zero cited papers in the dead category, the data yields the following average values: $m_L = 34.1$, $m_D = 4.5$ and $m = 12.8$. The fraction of live papers is $f = 27.0\%$. With an rms. error of only 21%, we can do a least squares fit of L_k to the distribution of live papers with parameters $k_0 = 65.6$, $a = 0.436$, and $b = 12.4$. Although only the live data (the squares in figure 2) is fitted, the agreement with the empirical data in figures 2 and 3 is quite striking.

From the model parameters k_0, a, b, we can calculate mean citation numbers for the fit of 32.9, 4.25, and 12.8 for the live, dead, and total population respectively; the fraction of live papers is found to be 29.8%. More interestingly, we learn from the fit that 7.5% of the papers with 0 citations *are actually alive*. If we assign this fraction of the zero-cited papers to the live population, we find the following corrected values for the average values 31.5, 4.6 and 12.5 for the live, dead, and total population respectively; the fraction of live papers

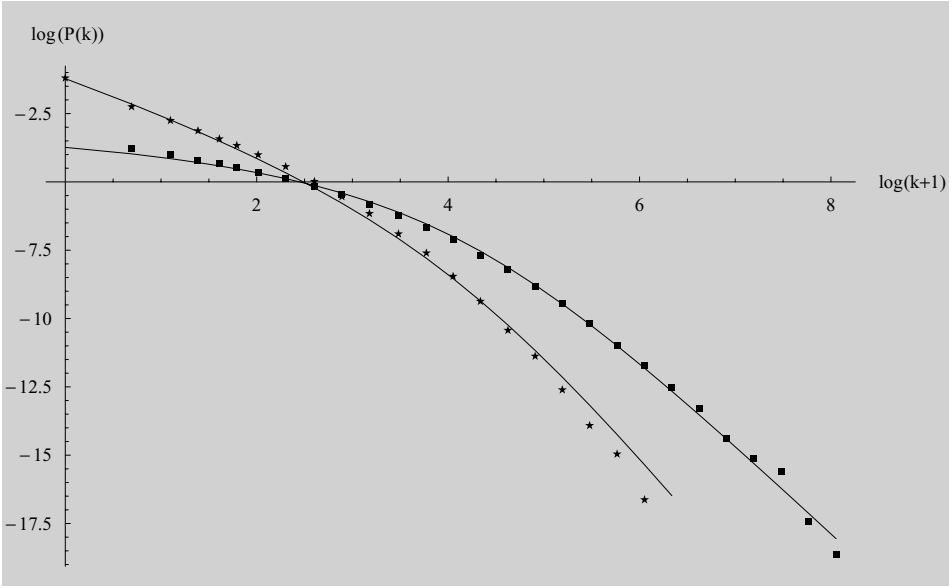


Figure 2. Log-log plots of the normalized degree distributions of live and dead papers. The filled squares represent the live data and the stars represent the dead data. Both lines are the result of a fit to the live data (filled squares) alone.

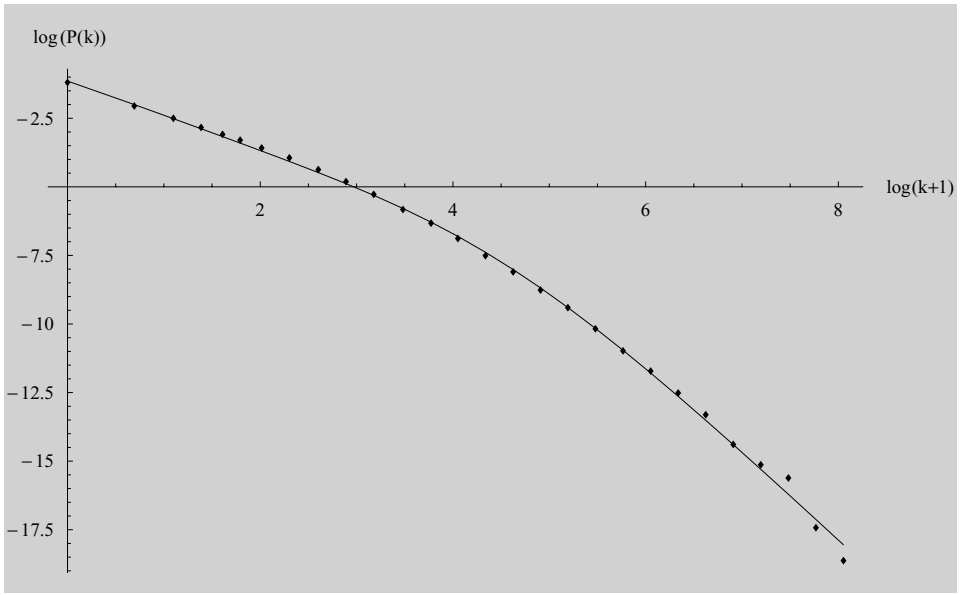


Figure 3. A log-log plot of the normalized degree distribution of all papers (live plus dead). The points are the data; the fit (solid line) is derived from the fit to the live papers (filled squares) in figure 2.

is adjusted to become 29.2%. Again, this is a striking agreement with the data. There is so little strain in the fit that we could have determined the model parameters from the empirical values of m_L, m_D, and f. Doing this yields only small changes in the model parameters and results in a description of comparable quality!

Figure 2 reveals that fitting to the live distributions, results in systematic errors for high values of k when we extend the fit to describe the dead papers, but this is not surprising. Recall the similarly systematic deviations from the straight line seen in figure 1. This figure also explains why the fit to the total distribution shows no deviations from the fit for high k-values even though the total fit includes both live and dead papers—live papers dominate the total distribution in this regime. The obvious way to fix this problem is via a small modification of the η_k. In summary, the full model is able to fit the distributions of both live and dead papers with remarkable accuracy.

One drawback, with regard to the full solution is the relatively impenetrable expression for L_k in Eq. (7)—associating any kind of intuition to the conglomerate of gamma-functions presented there can be difficult. Let us therefore demonstrate that L_k can be well approximated by a two power law structure. We begin by noting that, in the limit of large k_0 (as it is the case here), the values of k_1 and k_2 are simply

$$k_1 = -\frac{1}{a} + \frac{b}{ak_0} - k_0 \tag{15}$$

$$k_2 = -1 - \frac{b}{ak_0}. \tag{16}$$

Now, let us write out only the k-dependent terms in Eq. (7) and assign the remaining terms to a constant, C

$$L_k = C\frac{(k+k_0-1)!}{(k-k_1)!}\frac{(k+1)!}{(k-k_2)!} \tag{17}$$

$$\approx C\frac{1}{(k+k_0-1)^{1-k_0-k_1}}\frac{1}{(k+1)^{-(1+k_2)}} \tag{18}$$

$$\approx C\frac{1}{(k+k_0-1)^{1+\frac{1}{a}-\frac{b}{ak_0}}}\frac{1}{(k+1)^{\frac{b}{ak_0}}}, \tag{19}$$

In Eq. (18), we have utilized the fact that

$$\frac{(x+s)!}{x!} \approx x^s \tag{20}$$

when $x \to \infty$, and in Eq. (19) we have inserted the asymptotic forms of k_1 and k_2, from Eqs. (15) and (16).

This expression for L_k in Eq. (19) is only valid for large k and k_0, but it proves to be remarkably accurate even for smaller values of k. With the asymptotic forms of k_1 and k_2 inserted, we can explicitly see that the first power law is largely due to preferential attachment and that the second power law is exclusively due to the death mechanism. The form for very large k is unaltered by the parameter b. This is not surprising, since there is a low probability for highly cited papers to die. We see that the primary role of the death mechanism in the full model is to add a little extra structure to the L_k for small k.

7. Conclusions

Compelled by a significant inhomogeneity in the data, we have created a model that provides an excellent description of the SPIRES database. It is obvious that the death mechanism ($b \neq 0$) is essential for describing the live and dead populations separately, but less clear that it is indispensable when it comes to the total data. Fitting the total distribution with a preferential attachment only model ($b = 0$) results in $a = 0.528$ and $k_0 = 13.22$ and with a rms. fractional error of 33.6%. This fit displays systematic deviations from the data, but considering that the fit ignores important correlations in the dataset, the overall quality is rather high. The important lesson to learn from the work in this paper, is that even a high quality fit to the global network distributions is not necessarily an indication of the absence of additional correlations in the data.

The most significant difference between the full live-dead model and the model described above is expressed in the value of the parameter k_0. The value of this parameter changes by a factor of approximately 5, from 65.6 to 13.2. It strikes us as natural that preferential attachment will not be important until a paper is sufficiently visible for authors to cite it without reading it. We thus believe that $k_0 \approx 66$ is a more intuitively appealing value for the onset of preferential attachment. However, independent of which value of the k_0 parameter one prefers, the comparison of these two models clearly demonstrates the danger

of assigning physical meaning to even the most physically motivated parameters if a network contains unidentified correlations or if known correlations are neglected in the modeling process. Specifically, it would be ill advised to draw strong conclusions about the onset of preferential attachment if the death mechanism is not included in the model making.

In summary, the live and dead papers in the SPIRES database constitute distributions with significantly different statistical properties. We have constructed a model which includes modified preferential attachment and the death of nodes. This model is quantitatively successful in describing the citation distributions for live and dead papers. The resulting model has also been shown to produce a two power law structure. This structure provides an appealing link to the work in Lehmann et al. (2003), where a two power law structure was adopted to characterize the form of the SPIRES data without any theoretical support. Finally, we have been shown that even in the absence of preferential attachment, the death mechanism alone can result in power laws. Since many real world networks have a large number of inactive nodes and only a small fraction of active nodes, we are confident that this mechanism will find more general use.

Notes

1. More precisely, de Solla Price was the first person to re-think Simon's model and use it as a basis of description for *any* kind of network, cf. Newman (2003).
2. SPIRES is an acronym for 'Stanford Physics Information REtrieval System' and is the oldest computerized database in the world. The SPIRES staff has been cataloguing all significant papers in high energy physics and their lists of references since 1974. The database is open to the public and can be found at http://www.slac.stanford.edu/spires/.
3. For present purposes, this is appropriate when $r \geq 2$. When $r < 2$, the neglected factor is essential for ensuring the convergence of the average number of citations for the live and dead papers m_L and m_D.

References

Albert, R. and A.-L. Barabási (2002), "Statistical Mechanics of Complex Networks," *Reviews of Modern Physics*, 74, 47.

Barabási, A.-L. and R. Albert (1999), "Emergence of Scaling in Random Networks," *Science*, 286, 509.

de Solla Price, D. (1965), "Networks of Scientific Papers," *Science*, 149, 510–515.

de Solla Price, D. (1976), "A General Theory of Bibliometric and Other Cumulative Advantage Processes," *Journal of the American Society for Information Science*, 27, 292.

Dorogovtsev, S.N. and J.F.F. Mendes (2002), "Evolution of Networks," *Advances in Physics*, 51, 1079.

Klemm, K. and V.M. Eguíluz (2002), "Highly Clustered Scale-Free Networks," *Physical Review E*, 65, 036123.

Krapivsky, P.L. and S. Redner (2001), "Organization of Growing Random Networks," *Physical Review E*, 63, 066123.

Krapivsky, P.L., S. Redner, and F. Leyvraz (2000), "Connectivity of Growing Random Networks," *Physical Review Letters*, 85(21), 4629.

Lehmann, S., B.E. Lautrup, and A.D. Jackson (2003), "Citation Networks in High Energy Physics," *Physical Review E*, 68.

Lehmann, S., A.D. Jackson, and B.E. Lautrup (2005), "Life, Death, and Preferential Attachment," *Europhysics Letters*, 69, 298.

Newman, M.E.J. (2003), "The Structure and Function of Complex Networks," *SIAM Review*, 45, 167.

Redner, S. (1998). "How Popular is Your Paper? An Emperical Study of the Citation Distribution," *European Physics Journal B*, 4, 131–134.

Redner, S. (2004), "Citation Statistics From More Than a Century of Physical Review," *Physics/0407137*.
Simon, H.A. (1957), *Models of Man*. Wiley, New York.
Watts, D.J. (1999), *Small Worlds*. Princeton University Press, Princeton.

Sune Lehmann is pursuing his Ph.D. at the Technical University of Denmark. In 2003 he received his M.Sc. degree from the Niels Bohr Institute in Copenhagen, Denmark. Currently, his research are primarily the physics of complex systems, in particular complex networks in sociology and biophysics.

Andrew D. Jackson is professor of theoretical physics at the Niels Bohr Institute, Copenhagen, Denmark. His primary research activities are related to non-perturbative quantum field theories and the theory of strongly interacting quantum fluids including Bose-Einstein condensates. His work on complex systems, particularly that related to random matrix theory, is motivated by these primary interests. Jackson has an active interest in the history of science and has published books and articles on Hans Christian Orsted. He is a Fellow of the American Physical Society and a member of the Royal Danish Academy of Sciences and Letters.

Cited by

The paper *Live and Dead Nodes* has been cited by the following papers [50, 85].

CHAPTER 5

Bayesian Analysis

BASED on the fact that the Roman Empire's excellent road system radiated from the capital like the spokes of a wheel, the metaphor that "All roads lead to Rome" was already being used as early as the 1100s. In the spirit of this metaphor, we will now follow a completely new path towards an understanding of the structure of complex networks.

As a scientist, the project of uncovering the structure of complex networks, is an end in itself. The goal of science is to understand nature. Most analyses of networks do, however, have practical applications. Exploration of social networks lead to a better understanding of the spreading of epidemics, or enable us to navigate more intelligently in the social domain. Investigating the network of the power-grid enables us to understand the nature of the major black-outs that periodically affect densely populated areas in both Europe and America [4]. Knowledge of the topology of the P2P networks lets us design intelligent search- and download schemes [2, 3].

The network of scientific citations network enhances our knowledge of the structure of scientific knowledge; via the author-network it also illuminates the relationships between scientists. The main application derived from the study

of this network stems from the fact that the paper network can help us answer the emotionally charged question of which scientists are 'good' and which scientists are 'bad'. Specifically, it is common to assume that the number of citations of a scientific papers is an indicator of the quality of that paper. — If a paper has many citations, its content has been used by many other scientists; if a paper has only a few citations, this means that results in the paper have not been used in other people's work[1].

While it is relatively unproblematic to discuss the quality of a single paper, based on its number of citations, it is less obvious how one should judge the quality of a scientist based on a list of the papers that he/she has written. *How do we go from a scientist's list of citations to a scalar measure of his/her quality as a researcher?*

In Sections 5.1 and 5.2 Lehmann *et al.* set out to answer this question in a statistical manner. Although the starting point clearly concerns an application of the network data, the path to obtaining an answer has opened up an interesting approach towards analysis of more features than simply nodes and links. In the case of the complicated author-network shown in figure 3.2, we represent author (node) by a set of numbers, corresponding to the citations (in-degrees) of each of his papers. These citation records are analyzed using Bayesian statistics [39]. This approach does not take into account the network 'structure'—in the strict sense of the arrangement of links between nodes—but, as we shall see in the following—it does, for example, provide valuable information about groupings of authors (nodes) and the level of stratification in the network.

In chapter 1, we have discussed the perils of using classical statistical measures to describe complex networks. This criticism does not extend to a Bayesian framework, because the Bayesian scheme forces us to make our assumptions about the network explicit. *Qua* the SPIRES data base, we possess an extremely detailed knowledge about the complicated network of papers and authors; the Bayesian framework allows us utilize this information in a systematic way. A Bayesian approach can be a helpful tool in the quest for a deeper understanding of network structure, whenever we wish to include information about the network that extends beyond the simple adjacency matrix.

[1] There are many caveats to respect if one makes the assumption that citations are a proxy for quality. See Section 5.2 for an in-depth discussion of this subject.

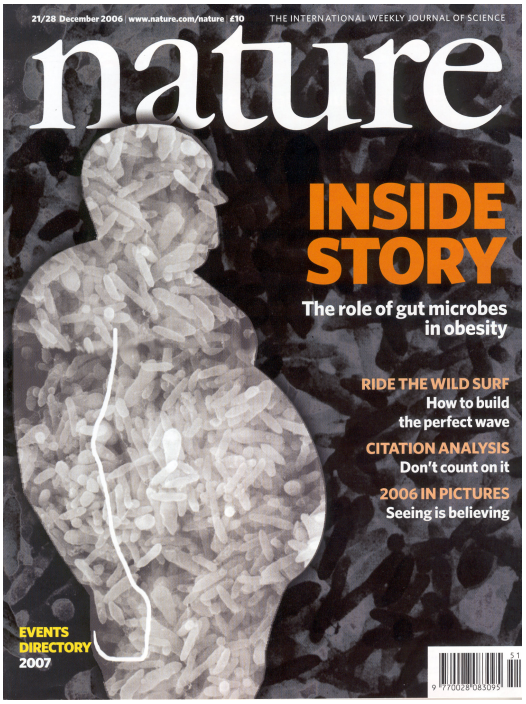


Figure 5.1: The front page of *Nature* on December 28th, 2006. The fact that our analysis of the network of scientific citations is mentioned on the front page, serves to illustrate the high level of personal involvement and interest that scientists exhibit, when when their research is evaluated. Since the recent publication of this paper, we have received hundreds of emails, been the subject of several newspaper articles, participated in radio-shows, and started to become a visible part of blogosphere with appearances in dozens of science-blogs.

5.1 Measures for Measures

The paper *Measures for Measures*[2] [56] which reports some of the results obtained using the Bayesian analysis was published as a commentary in the Journal *Nature*. Due to the commentary format, the focus is primarily on the application side of the network of scientific citations. In this paper, Bayesian statistics is applied in order to investigate the reliability of several well known indicators of scientific quality. Specifically, the *number of papers published per year*, *the mean number of citations per paper*, and a new (but popular) measure called the *Hirsch index* [37] were investigated. A scientist is said to have Hirsch index h if h of their total, N, papers have at least h citations each, and the remaining $(N - h)$ papers have fewer than h citations. Lehmann *et al.* demonstrate that the mean number of citations is the best measure. The h-index turns out (somewhat surprisingly) to be a poor measure of scientific quality. Measuring the number of papers published per year is as ineffective as assigning quality to authors based

[2] Aside from the literal meaning of the title of this paper, the title refers to a play called *Measure for Measure*, by William Shakespeare [86].

on their first initial.

At the *Nature* website, substantial supplementary online information accompanies the main text. Here, the supplementary information is presented directly after the main text.

COMMENTARY

Measures for measures

Are some ways of measuring scientific quality better than others? **Sune Lehmann**, **Andrew D. Jackson** and **Benny E. Lautrup** analyse the reliability of commonly used methods for comparing citation records.

Although quantifying the quality of individual scientists is difficult, the general view is that it is better to publish more than less and that the citation count of a paper (relative to citation habits in its field) is a useful measure of its quality. How citation counts are weighed and analysed in practice becomes important as publication records are increasingly used in funding, appointment and promotion decisions. Typically, a scientist's full citation record is summarized by simpler measures, such as average citations per paper, or the recently proposed Hirsch index[1], which is ever more being used as an indicator of scientific quality[2]. Despite their growing importance, there have been few attempts to discover which of the popular citation measures is best and whether any such measure is statistically reliable.

"There have been few attempts to discover which of the popular citation measures is best and whether any are statistically reliable."

Measures of citation quality are of value only if they can be assigned to individual authors with high confidence. Previous bibliometric studies[2] have compared different measures of scientific quality, but just because two measures agree does not mean that either one is accurate or reliable. We will argue that some citation-based measures can provide useful information given data of sufficient quality, but others fail to meet minimum acceptable standards. This should concern every working scientist.

Unfair discrimination

Because citation practice differs markedly between disciplines and subfields, a homogeneous set of authors is essential for any statistical analysis of citations. Here we use data from the theory section of the SPIRES database in high-energy physics, which has the requisite homogeneity[3]. Within this database, the probability that a paper will receive k citations falls slowly with increasing k and is described by a power-law distribution, a/k^{γ} with $\gamma \approx 2.8$, for large k. This long-tailed distribution has a number of consequences. About 50% of all papers have two or fewer citations; the average number of citations is 12.6. The top 4.3% of papers produces 50% of all citations whereas the bottom 50% of papers yields just 2.1% of all citations. Measuring an author's mean or median citation count per paper probe different aspects of their full citation record: which is better? Fortunately, this question can be posed in a way that yields a statistically compelling answer.

The purpose of comparing citation records is to discriminate between scientists. An author's citation record is a list of the number of citations of each of the author's publications. Until reduced to a single number, this list cannot provide a means of ranking scientists. But whatever the intrinsic merits of the chosen number, it will be of no practical use unless the uncertainty in assigning it to individual scientists is small. From this perspective, the 'best' measure will be that which minimizes uncertainty in the values assigned and hence maximizes discrimination between individuals. We analyse three measures of author quality: mean number of citations per paper, number of papers published per year, and the Hirsch index. A scientist is said to have Hirsch index h if h of their total, N, papers have at least h citations each, and the remaining $(N\text{-}h)$ papers have fewer than h citations[1]. For this study, we adopt Hirsch's assumption that h divided by N "should provide a useful yardstick". To calibrate our results, we also consider an obviously meaningless measure; we rank authors alphabetically by name.

We start with one of the three measures we had chosen and sort the SPIRES authors into decile bins. We use the full citation records for all authors in a given bin, n, to calculate the conditional probability that a paper written by an author in bin n will have k citations. From these conditional probabilities, we use Bayes' theorem to determine the average probability that an author initially assigned to bin n should instead be assigned to bin m. (To do this, we calculate the probability that the full publication record of each author in bin n was drawn, at random, on the conditional probability appropriate for bin m; see Supplementary Information.) Because the m assignment is based on an author's full citation record, it is more reliable than the n assignment. This process is repeated for each decile bin and for each measure considered.

Quality testing

A perfect measure of author quality would place all weight in the diagonal entries of a plot of m versus n (Fig. 1, overleaf). The better the measure, the more weight will be found in the diagonal boxes. Figure 1 reveals that both accuracy and certainty are sensitive to the choice of indicator.

An alphabetical ranking of authors contains no information regarding scientific quality, and so every author is assigned to every decile with equal probability (Fig. 1a). The resulting root-mean-square (*rms*) uncertainty in author assignment thus has the maximum value of ±29 percentile points. One of the most widely used measures of scientific quality is the average number of papers published by an author per

A. MACDONALD

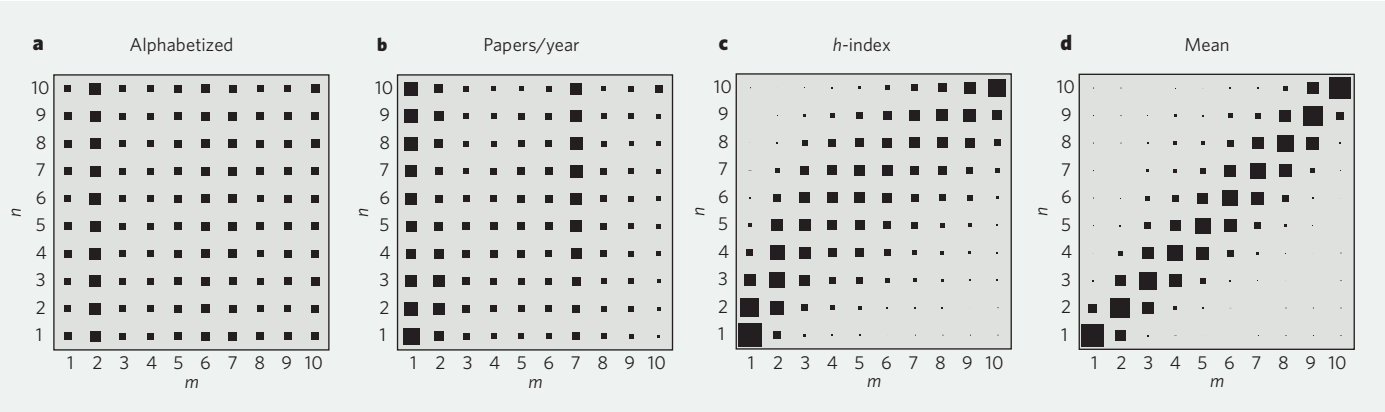


Figure 1 | The probabilities for four different measures. a–d, Each horizontal row, indexed by n, shows the average probabilities that authors initially assigned to a given decile bin n are predicted to lie in a different decile bin m. The probabilities are proportional to the areas of the corresponding squares.

year[1,4]. This measure has a similar *rms* variation to alphabetization (Fig. 1b). Publication frequency would be more useful if all papers were cited equally but, as noted above, this is not the case. The best that can be said of publication frequency is that it measures industry rather than ability.

Impact factors are widely used to introduce a citation measure into calculations of publication frequency. The impact factor for each journal, as defined by Thomson Scientific/ISI[5], is the average number of citations acquired during the past two years for papers published over the same period. But weighting each paper by the journal's current impact factor is unlikely to improve the situation, especially when estimating scientific quality across an author's entire career. The impact factor for reputable journals is determined by a small fraction of highly cited papers, so the citation rate for individual papers is largely uncorrelated to the impact factor of the journal in which it was published[6]. The widespread use of publication frequency — with or without an impact factor — is disturbing and requires further study.

Word of caution

Hirsch's h-index attempts to strike a balance between productivity and quality and to avoid the heavy weight that power-law distributions place on a relatively small number of highly cited papers. Hirsch's measure is obtained by ranking papers in order of decreasing citations with paper i having $C(i)$ citations and solving the equation $h = C(h)$. This is the simplest version of $h = AC(h)^K$. Hirsch's choice of $A = K = 1$ is unsupported by any data. Nevertheless, Fig. 1 indicates that this measure does better than publication frequency, because the h-index depends on the entire citation record.

Hirsch's measure overestimates the initial n-assignments by some 8 percentile points as indicated by higher densities above the diagonal (Fig. 1c). Moreover, the *rms* uncertainty in the assignment of h is ±16 percentile points, which is only a factor of two better than alphabetization. Although capturing certain aspects of quality, Hirsch's index cannot make decile assignments with confidence.

Compared with the h-index, the mean number of citations per paper is a superior indicator of scientific quality, in terms of both accuracy and precision. The average assignment of each n-bin is in error by 1.8 percentile points with an associated *rms* uncertainty of ±9. Similar calculations based on authors' median citation give an accuracy of 1.5 and an uncertainty of only ±7 percentile points, suggesting that the median copes better with long-tailed distributions.

Simple scaling arguments[4] show that the *rms* uncertainty for any measure decreases rapidly (exponentially) as the total number of papers increases. Thus, for example, no more than 50 papers are required to assign a typical author to deciles 2–3 or 8–9 with 90% confidence when using the mean citation rate as a measure. Fewer papers suffice for deciles 1 and 10. Any attempt to assess the quality of authors using substantially fewer publications must be treated with caution.

Data access

The methods used here are not specific to high-energy physics. Given suitably homogeneous data sets, they can be applied to any scientific field and permit a meaningful (probabilistic) comparison of scientists working in different fields by assuming the equality of scientists in the same percentile of their respective peer groups. Similarly, probabilities can be combined to make meaningful quality assignments to authors with publications in several disjointed subfields.

There are strong indications that an author's initial publications are drawn on the same probability distribution as their remaining papers[7]. Therefore, with sufficient numbers of publications to draw meaningful conclusions (50 or more) the mean or median citation counts can be a useful factor in the academic appointment process.

Unfortunately, the potential benefits of careful citation analyses are overshadowed by their harmful misuse. Institutions have a misguided sense of the fairness of decisions reached by algorithm, and unable to measure what they want to maximize (quality), institutions will maximize what they can measure. Decisions will continue to be made using measures of quality that either ignore citation data entirely (such as frequency of publication) or rely on data sets of insufficient quality.

"Institutions have a misguided sense of the fairness of decisions reached by algorithm; unable to measure what they want to maximize (quality), they will maximize what they can measure."

Access to the full citation distribution for an entire subfield is essential to our analysis. Existing databases such as the ISI can therefore actively help to improve the situation by compiling field-specific homogeneous data sets similar to what we have generated for SPIRES. This would allow institutions and scientists alike to evaluate the quality of any citation record using all available information. For their part, scientists should insist that their institutions disclose their uses of citation data, making both data and the methods used for data analysis available for scrutiny. In the meantime, we shall have to continue to do things the old-fashioned way and actually read the papers. ■

Sune Lehmann is at the Department of Informatics and Mathematical Modeling, Technical University of Denmark, DK-2800, Lyngby, Denmark.
Andrew D. Jackson and Benny E. Lautrup are at The Niels Bohr Institute, Blegdamsvej 17, DK-2100, Copenhagen, Denmark.

1. Hirsch, J. E. *Proc. Natl Acad. Sci. USA* **102,** 16569 (2005).
2. van Raan, A. F. J. *Scientometrics* **67,** 491 (2006).
3. Lehmann, S., Lautrup, B. E. & Jackson, A. D. *Phys. Rev. E* **68,** 026113 (2003).
4. van Raan, A. F. J. *J. Am. Soc. Inf. Sci.* **57,** 408 (2005).
5. Thomson Scientific/ISI http://www.isinet.com/
6. Seglen, P. O. *J. Am. Soc. Inf. Sci.* **45,** 1 (1994).
7. Lehmann, S. Thesis, The Niels Bohr Institute (2003).

Supplementary information accompanies this Commentary on *Nature*'s website.

Measures for Measures

Supplementary Online Information

S. Lehmann[1]
A. D. Jackson
B. E. Lautrup

[1]Electronic Address: lehmann@nbi.dk

Contents

1 Data

1.1 Acquisition

This section provides a short description of the acquisition and processing of data from the SPIRES (Stanford Public Information REtrieval System) data base. Ultimo 2003, the database manager[1] provided us with a text file containing the following information for each paper in spires: Title, List of authors, Publication information, References, Sub-field classification, and Keywords. We added this information to a relational data base (MySQL) in order to create a network of authors and papers. The data used here was generated by querying the resulting data base. Thus, only citations from within the data base are counted. In order to ensure the validity of the data, we have also used an independent route to generate the data; we employed the programming language Perl to extract the relevant information from the main text file.

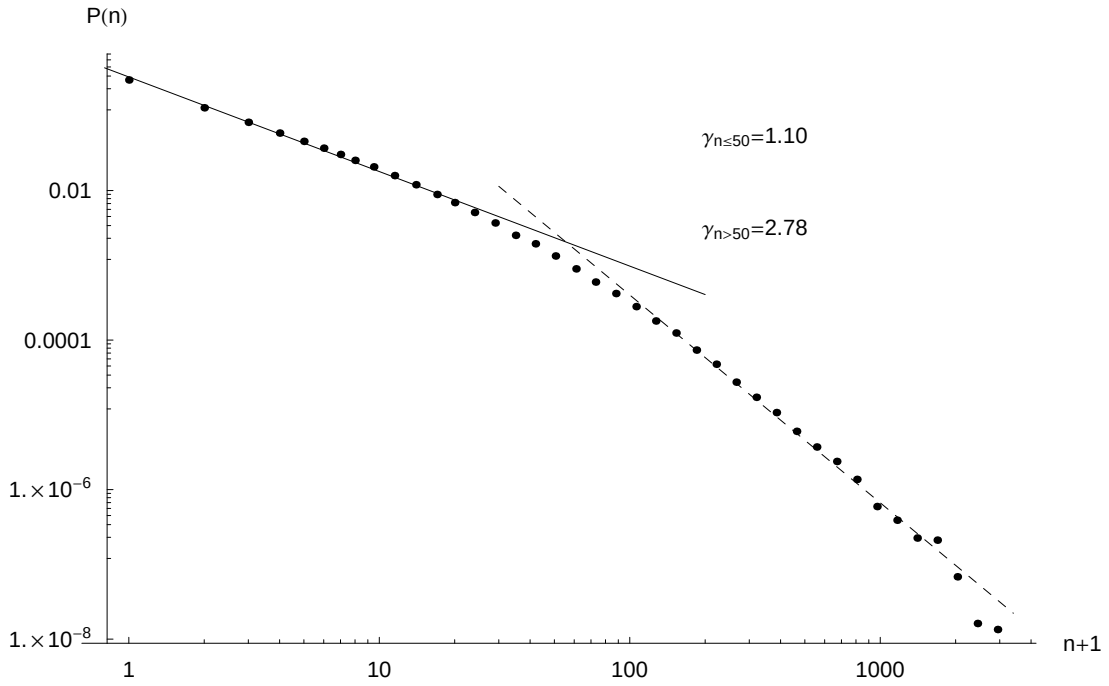
One main problem in processing this data is identifying authors uniquely, since the same author can represent his name in may different ways (e.g. John James Smith, John J. Smith, J. J. Smith, J. Smith, etc.). For the data shown, authors were identified by last name and first two initials. Checks were performed using (i) last name and all initials and (ii) last name and first initial only. These two cases represent approximate upper and lower bounds on the number of unique authors in the data base. No significant changes were found in either case.

1.2 Statistics

Our data set consists of all publications by "academic scientists"—defined as those with 25 or more published papers—in the theory subfield of SPIRES. The resulting data set contains 274 470 papers written by 6 737 authors; this data set is highly homogeneous [1]. One possible description of the distribution of citations of papers is a double power-law structure[2]. Specifically the probability that a paper will recieve n citations is approximately proportional to $(n+1)^{-\gamma}$ with $\gamma = 1.10$ for $n \leq 50$ and $\gamma = 2.78$ for $n > 50$. These features of the global distribution are also present in the conditional probabilities for subgroups of authors binned according to most measures of quality. In virtually all cases, the conditional probabilities can also be described accurately by separate power-laws in each of two regions with a relatively sharp transition between the regions. As one might expect, authors with more citations are described by flatter distributions (i.e., smaller values of γ) and a somewhat higher transition point. Supplementary Figure 1 displays the total distribution of citations as a binned and normalized histogram.

[1] Travis C. Brooks from the SLAC Library.

[2] The double power-law description is only one of many possible parameterizations of the data; better fits to the data can certainly be made, but any increase in the number of parameters demands a justification.



Supplementary Figure 1: Logarithmically binned histogram of the citations counts of all papers by authors with more than 25 publications in the theory subsection of SPIRES. The data is normalized and the axes are logarithmic.

2 The Bayesian Method

We have binned the SPIRES authors and their citation records according to each of the four tentative measures, m, described in the main paper. Studies performed on the first 25, first 50 and all papers of authors with a given value of m indicate the absence of temporal correlations in the citation distributions of individual authors. In practice, we bin authors in deciles according to their value of m and papers logarithmically, due to the asymptotic power law behavior noted above. We have confirmed that the results here are relatively insensitive to binning effects.

We have constructed the prior distribution, $p(\alpha)$, that an author is in author bin α (in the case of decile bins $p(\alpha) = 1/10$ for all bins) and the conditional probability, $P(i|\alpha)$, that a paper by an author in bin α will fall in citation bin i. For each bin α, the $P(i|\alpha)$'s are simply citation distributions analogous to the normalized histogram displayed in Supplementary Figure 1, but constructed using only papers written by authors in bin α.

Now, we wish to calculate the probability, $P(\{n_i\}|\alpha)$, that an author in bin α will have a citation record with n_i papers in each citation bin i. To do this, we assume[3] that citations for the M papers written by a given author with n_i papers in citation bin i are obtained

[3]The argument here is based on the additional simplifying assumption that the distribution of total papers per author is the same in all author bins. This assumption, which is readily relaxed, has no significant effect on the results presented here.

from M independent random draws on the appropriate distribution, $P(i|\alpha)$. Thus,

$$P(\{n_i\}|\alpha) = M! \prod_i \frac{P(i|\alpha)^{n_i}}{(n_i)!} \ . \tag{1}$$

We have already noted the absence of large-scale temporal variations in $P(i|\alpha)$ during an author's scientific life. Other correlations could be present. For example, one particularly well-cited paper could lead to an increased probability of high citations for its immediate successor(s). While it is difficult to demonstrate the presence or absence of such correlations, the results below provide *a posteriori* indications that such correlations, if present, are not overly important. We can invert the probability $P(\{n_i\}|\alpha)$ using Bayes' Theorem to obtain

$$\begin{aligned} P(\alpha|\{n_i\}) &= \frac{P(\{n_i\}|\alpha)\, p(\alpha)}{p(\{n_i\})} \\ &= \frac{p(\alpha) \prod_k P(k|\alpha)^{n_k}}{\sum_{\alpha'} p(\alpha') \prod_{k'} P(k'|\alpha')^{n_{k'}}} \ . \end{aligned} \tag{2}$$

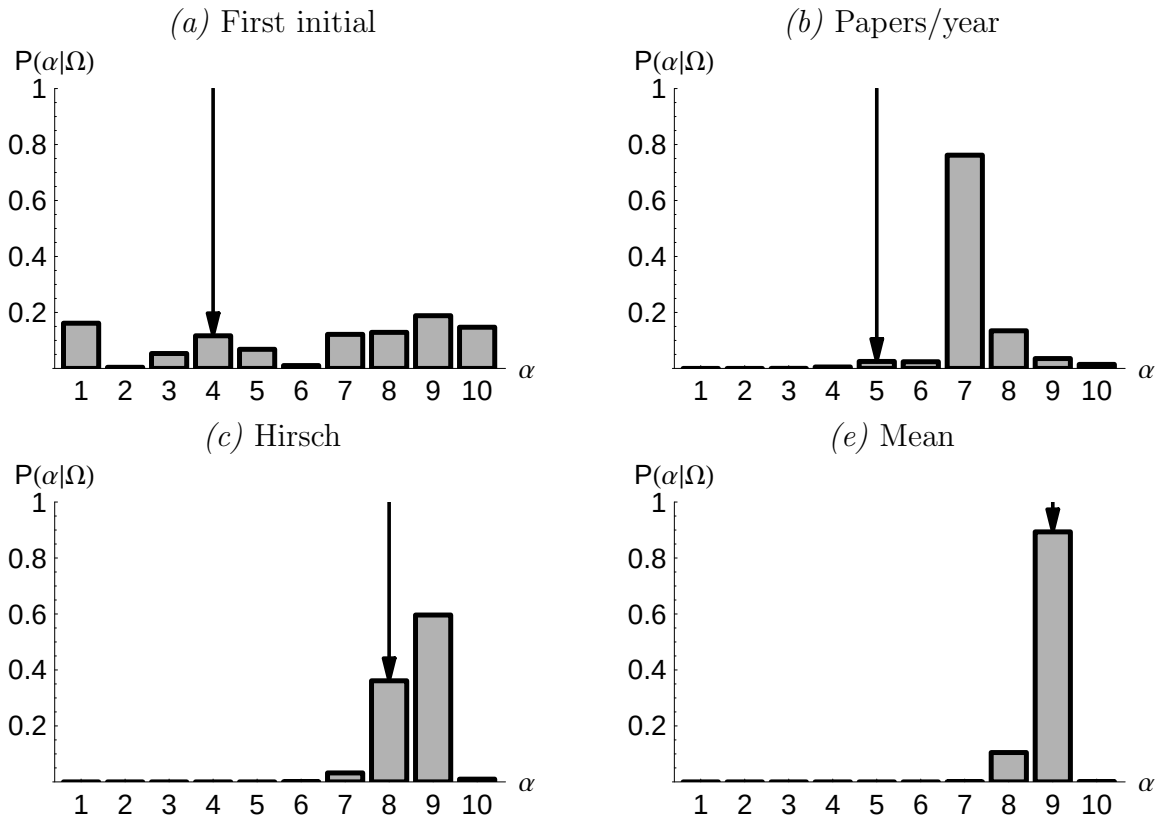
Note that the combinatoric factors cancel.

The quantity $P(\alpha|\{n_i\})$, which represents the probability that an author with citation record $\{n_i\}$ belongs in quality bin (i.e., decile) α, is of primary interest. While any given measure (e.g., the mean number of citations per paper) can be calculated immediately from an author's citation record $\{n_i\}$, the calculated values of $P(\alpha|\{n_i\})$ provide more detailed and reliable information. By exploiting differences between the various conditional probabilities, $P(\{n_i\}|\alpha)$, as a function of α, Supplementary Equation (2) determines the appropriate decile value of m (or its most probable value) using all statistical information in the data. By using the an author's full citation record, the large fluctuations which are inevitable in e.g. the number of citations of the author's maximally cited paper are thereby materially reduced. Further, by providing us with values of $P(\alpha|\{n_i\})$ for all α, we have a statistically trustworthy gauge of whether the resulting uncertainties in the assigned value of m are sufficiently small for it to be a reliable measure of author quality.

2.1 A Single Author Example

In short, Supplementary Equation (2) provides us with a measure of an author's expected lifetime quality along with information which allows us to assess the reliability of this determination. The confidence with which we can assign a value of m approaches 100% exponentially with the total number of published papers. As we shall see, it is also sensitive to the quality measure chosen. To gain an understanding of $P(\alpha|\{n_i\})$, let us consider a concrete example.

We will investigate the (real) citation record of author A with citation record Ω. Supplementary Figure 2 shows the probabilities that A will lie in each of the deciles using the four different measures defined in the main text. It is clear from the figure that there are significant differences in the results obtained, both in the apparent accuracy of their



Supplementary Figure 2: A single author example. We analyze the citation record of author A with respect to four different measures. Author A has written a total of 88 papers. The mean number of citations per paper is 26, Hirsch's h-index is 29 for this author, the maximally cited paper has 187 citations, and papers have been published at the average rate of 2.5 papers per year. The various panels give the probability that author A belongs to each of the ten deciles based on the corresponding measure; the vertical arrow shows the decile bin to which author A is assigned by direct calculation of each measure.

predictions and, more importantly, in the corresponding uncertainties. In all cases, large uncertainties are due to the fact that the conditional probabilities, $P(i|\alpha)$ are largely independent of α. Such independence is to be expected in the case of the alphabetic binning of authors, and the inability of the citation record to identify the first initial of author A's name is hardly surprising. The figure also suggests that, although this distribution has a peak, the number of papers published per year is unable to determine to which bin author A was assigned. The mean number of citations per paper provides an accurate determination with a small uncertainty, thus the use of Supplementary Equation (2) has compensated for the large fluctuations which might have been expected from the use of mean citation rate as a measure of quality. Hirsch's measure falls somewhere between the best and worst choice of measures.

2.2 Construction of Figure 1 in Main Paper

Measures of quality are of value only to the extent that they can be assigned to individual authors with high confidence. The methods described above allow us to determine this confidence for any choice of measure in a manner which is value-free and completely quantitative. In order to perform this evaluation, we repeat the calculations leading to Supplementary Figure 2 for all authors in the SPIRES database. We calculate the probability, $P(\beta|\alpha)$, which is the probability, averaged over the authors in author bin α, that the full citation record of an author initially assigned to bin α by the measure under consideration was drawn at random on the distribution $P(i|\beta)$, appropriate for author bin β. Stated simply, $P(\beta|\alpha)$ is the probability that an author assigned to be in bin α is predicted to lie in bin β. Thus, $P(\beta|\alpha)$ is the average

$$P(\beta|\alpha) = \frac{1}{N_\alpha} \sum_{\{n_i\}\in\alpha} P(\beta|\{n_i\}), \tag{3}$$

where N_α is the number of authors in bin α. The figure in the main paper is simply the "stacked" results of this calculation, that is, for each measure, we plot the array of probabilities

$$\begin{array}{cccc} P(1|10) & P(2|10) & & P(10|10) \\ \vdots & & \iddots & \\ P(1|2) & P(2|2) & & P(10|2) \\ P(1|1) & P(2|1) & \cdots & P(10|1) \end{array}, \tag{4}$$

where each probability $P(\beta|\alpha)$ is represented as a black square with area proportional to the corresponding probability.

2.3 Scaling

In this section, we will consider the question of how many published papers are required in order to make a reliable prediction of the lifetime quality measure for a given author. (Here, we will consider only results using the mean citation rate as a measure.) Obviously, if this number is sufficiently small, analysis along the lines presented here can provide a practical tool of potential value in predicting long-term scientific accomplishment. In order to address this question, we will look at how $P(m|\{n_i\})$ scales as a function of the the number of papers in each bin for an average author. Assume that an average author belonging to bin α draws M papers at random from the distribution of $P(n|\alpha)$. The most most probable number of papers in each citation bin will thus be given as $n_i = MP(i|\alpha)$. Inserting this result into Supplementary Equation (2) and discarding all fixed factors, we find that

$$P(\alpha|\{n_i\}) \sim p(\alpha) \left(\prod_i P(i|\alpha)^{P(i|\alpha)} \right)^M . \tag{5}$$

For the same citation record, $\{n_i\}$, a similar expression permits determination of the probability that this average author will be assigned to any bin. It is clear from Supplementary

Equation (5) that the probability of assigning this average author to the wrong bin will ultimately vanish exponentially with M. Given enough papers, the bin with the largest probability will ultimately dominate. To correctly assign the most probable to outer deciles 1, 2, 3 and 8, 9, 10 at the 90% confidence level requires respectively $M = 10$, 40, 50, and 50, 50, 30 papers.

All quality measures have difficulty in making correct assignments to deciles 4–7. This apparent difficulty is due to our decision to group authors by deciles. It can be understood by assuming that the distribution of intrinsic author quality has a maximum at some non-zero value. Such an assumption seems reasonable if we imagine that there is a natural high-end cutoff and that academic appointment procedures filter out the least able. For any such distribution, the probability density will be highest for authors in the vicinity of this maximum. The binning of authors by deciles or percentiles then invites us to make distinctions where no material quality difference exists. The results of the main figure in the actual commentary remind us that we cannot do so. On the other hand, the probability that an author can be correctly assigned to the bins $4, 5, 6, 7$ collectively on the basis of 50 publications is higher than 90%.

3 The Median

Here, we show that the median of $\mathcal{N} = (2N + 1)$ random draws on *any* normalized probability distribution, $q(x)$, is normally distributed in the limit $\mathcal{N} \to \infty$. To this end we define the integral of $q(x)$ as

$$Q(x) = \int^x q(x')dx' \tag{6}$$

Evidently, $Q(x)$ grows monotonically from 0 to 1 independent of $q(x)$. The 'median' of this sample is defined as that value of x such that *(i)* one draw has the value x, *(ii)* N draws have a value less than or equal to x, and *(iii)* N draws have a value greater than or equal to x. The probability that the median is at x is now given as

$$P_{x_{1/2}}(x) = \frac{(2N+1)!}{1!N!N!} q(x) Q(x)^N [1 - Q(x)]^N \ . \tag{7}$$

For large N, the maximum of $P_{x_{1/2}}(x)$ occurs at $x = x_{1/2}$ where $Q(x_{1/2}) = 1/2$. Expanding the logarithm of $P_{x_{1/2}}(x)$ about its maximum value, we see that

$$P_{x_{1/2}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x - x_{1/2})^2}{2\sigma^2}] , \qquad \sigma^2 = \frac{1}{4\mathcal{N} q(x_{1/2})^2} \ . \tag{8}$$

An identical argument applies for any percentile—not just the median. E.g., for constructing the distribution of the 90th percentile, we would construct the the probability that $9N$ draws have a value less than x, N draws have a value greater than x, and one draw has the value of x. The distribution of *any* percentile, $0 \leq z \leq 1$ measured with $\mathcal{N}$ random draws on any distribution is a Gaussian with a maximum at some x_z such that $Q(x_z) = z$ and $\sigma^2 \sim (\mathcal{N} q(x_z)^2)^{-1}$.

4 Explicit $P(\beta|\alpha)$

In this section we attach the actual probabilities behind the figure in the main text; the numbers below correspond to the array in Supplementary Equation 4. As a visual help, the diagonals are set in bold face.

4.1 First Initial

$$\begin{pmatrix}
0.0761 & 0.2104 & 0.0686 & 0.0709 & 0.0819 & 0.1148 & 0.1010 & 0.0747 & 0.0801 & \mathbf{0.1216} \\
0.0839 & 0.1869 & 0.0710 & 0.0772 & 0.0866 & 0.1100 & 0.1107 & 0.0818 & \mathbf{0.0876} & 0.1042 \\
0.0820 & 0.1902 & 0.0700 & 0.0760 & 0.0857 & 0.1071 & 0.1147 & \mathbf{0.0820} & 0.0868 & 0.1054 \\
0.0851 & 0.1698 & 0.0715 & 0.0781 & 0.0927 & 0.1080 & \mathbf{0.1248} & 0.0841 & 0.0887 & 0.0972 \\
0.0790 & 0.2127 & 0.0695 & 0.0736 & 0.0847 & \mathbf{0.1142} & 0.1113 & 0.0776 & 0.0817 & 0.0958 \\
0.0814 & 0.1886 & 0.0713 & 0.0757 & \mathbf{0.0887} & 0.1099 & 0.1144 & 0.0817 & 0.0857 & 0.1025 \\
0.0814 & 0.1986 & 0.0680 & \mathbf{0.0751} & 0.0851 & 0.1057 & 0.1154 & 0.0802 & 0.0858 & 0.1048 \\
0.0791 & 0.2029 & \mathbf{0.0719} & 0.0728 & 0.0831 & 0.1096 & 0.1052 & 0.0779 & 0.0826 & 0.1150 \\
0.0776 & \mathbf{0.2276} & 0.0703 & 0.0724 & 0.0822 & 0.1161 & 0.1028 & 0.0757 & 0.0800 & 0.0953 \\
\mathbf{0.0841} & 0.1885 & 0.0712 & 0.0770 & 0.0857 & 0.1089 & 0.1129 & 0.0816 & 0.0876 & 0.1025
\end{pmatrix}.$$

4.2 Papers Per Year

$$\begin{pmatrix}
0.4493 & 0.0979 & 0.0347 & 0.0319 & 0.0462 & 0.0415 & 0.2412 & 0.0276 & 0.0169 & \mathbf{0.0128} \\
0.3591 & 0.1180 & 0.0452 & 0.0453 & 0.0637 & 0.0565 & 0.2204 & 0.0437 & \mathbf{0.0273} & 0.0208 \\
0.3134 & 0.1118 & 0.0484 & 0.0503 & 0.0674 & 0.0614 & 0.2388 & \mathbf{0.0536} & 0.0320 & 0.0228 \\
0.2321 & 0.1018 & 0.0518 & 0.0616 & 0.0839 & 0.0758 & \mathbf{0.2547} & 0.0683 & 0.0407 & 0.0292 \\
0.2321 & 0.1280 & 0.0672 & 0.0674 & 0.0861 & \mathbf{0.0780} & 0.1994 & 0.0649 & 0.0436 & 0.0332 \\
0.2130 & 0.1256 & 0.0679 & 0.0711 & \mathbf{0.0891} & 0.0792 & 0.2051 & 0.0699 & 0.0455 & 0.0336 \\
0.2024 & 0.1308 & 0.0768 & \mathbf{0.0746} & 0.0885 & 0.0811 & 0.1855 & 0.0734 & 0.0492 & 0.0378 \\
0.2747 & 0.1563 & \mathbf{0.0805} & 0.0665 & 0.0750 & 0.0692 & 0.1335 & 0.0621 & 0.0452 & 0.0369 \\
0.3077 & \mathbf{0.1741} & 0.0852 & 0.0642 & 0.0699 & 0.0661 & 0.0946 & 0.0529 & 0.0465 & 0.0388 \\
\mathbf{0.3406} & 0.1751 & 0.0841 & 0.0576 & 0.0590 & 0.0573 & 0.0774 & 0.0538 & 0.0482 & 0.0469
\end{pmatrix}.$$

4.3 Hirsch

$$\begin{pmatrix}
0.0000 & 0.0000 & 0.0010 & 0.0051 & 0.0124 & 0.0375 & 0.0805 & 0.1457 & 0.2298 & \mathbf{0.4881} \\
0.0000 & 0.0004 & 0.0105 & 0.0325 & 0.0593 & 0.1145 & 0.1703 & 0.2169 & \mathbf{0.2205} & 0.1752 \\
0.0000 & 0.0048 & 0.0503 & 0.0930 & 0.1292 & 0.1585 & 0.1671 & \mathbf{0.1671} & 0.1498 & 0.0801 \\
0.0003 & 0.0277 & 0.1150 & 0.1541 & 0.1789 & 0.1658 & \mathbf{0.1294} & 0.1041 & 0.0811 & 0.0435 \\
0.0046 & 0.0945 & 0.1787 & 0.1747 & 0.1745 & \mathbf{0.1413} & 0.1011 & 0.0704 & 0.0459 & 0.0142 \\
0.0248 & 0.2102 & 0.2253 & 0.1682 & \mathbf{0.1499} & 0.0957 & 0.0605 & 0.0405 & 0.0190 & 0.0059 \\
0.0711 & 0.3251 & 0.2157 & \mathbf{0.1322} & 0.1118 & 0.0665 & 0.0356 & 0.0211 & 0.0181 & 0.0027 \\
0.2243 & 0.4026 & \mathbf{0.1656} & 0.0768 & 0.0592 & 0.0352 & 0.0195 & 0.0101 & 0.0038 & 0.0029 \\
0.5417 & \mathbf{0.3180} & 0.0761 & 0.0315 & 0.0196 & 0.0071 & 0.0030 & 0.0030 & 0.0000 & 0.0000 \\
\mathbf{0.8844} & 0.0981 & 0.0104 & 0.0039 & 0.0032 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000
\end{pmatrix}.$$

4.4 Mean

$$\begin{pmatrix}
0.0000 & 0.0000 & 0.0000 & 0.0005 & 0.0000 & 0.0039 & 0.0049 & 0.0253 & 0.2087 & \mathbf{0.7567} \\
0.0000 & 0.0000 & 0.0006 & 0.0081 & 0.0038 & 0.0337 & 0.0493 & 0.2089 & \mathbf{0.6062} & 0.0895 \\
0.0000 & 0.0000 & 0.0015 & 0.0157 & 0.0185 & 0.0747 & 0.2037 & \mathbf{0.4388} & 0.2434 & 0.0036 \\
0.0000 & 0.0000 & 0.0104 & 0.0224 & 0.0563 & 0.2039 & \mathbf{0.4086} & 0.2566 & 0.0414 & 0.0003 \\
0.0000 & 0.0005 & 0.0257 & 0.0656 & 0.1873 & \mathbf{0.3843} & 0.2648 & 0.0654 & 0.0063 & 0.0000 \\
0.0000 & 0.0026 & 0.0619 & 0.1915 & \mathbf{0.4041} & 0.2600 & 0.0697 & 0.0096 & 0.0005 & 0.0000 \\
0.0000 & 0.0322 & 0.2127 & \mathbf{0.4104} & 0.2706 & 0.0646 & 0.0086 & 0.0007 & 0.0000 & 0.0000 \\
0.0028 & 0.1826 & \mathbf{0.5034} & 0.2542 & 0.0505 & 0.0060 & 0.0004 & 0.0000 & 0.0000 & 0.0000 \\
0.1037 & \mathbf{0.6462} & 0.2212 & 0.0266 & 0.0022 & 0.0001 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
\mathbf{0.8044} & 0.1882 & 0.0071 & 0.0002 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000
\end{pmatrix}.$$

References

[1] Lehmann, S., Lautrup, B. E., and Jackson, A. D. *Physical Review E* **68**, 026113 (2003).

5.2 A Quantitative Analysis of Measures of Quality

The paper *A quantitative analysis of measures of quality in science* [57], is the 'engine' behind the commentary [56]. Here, the comprehensive set of arguments and calculations that comprise the foundation of the conclusions in [56], is presented.

In particular, the new data set from SPIRES is described in great detail, four additional measures of quality are presented and analyzed, the effects of binning the data into deciles are analyzed, and the role of the Kullback-Leibler divergence between the conditional citation distributions is explored fully and used to illuminate the scaling arguments.

This paper has recently been submitted to the journal *Physical Review E* and is currently available from the online preprint server http://arXiv.org/ under the label *physics/0701311*.

A Quantitative Analysis of Measures of Quality in Science

Sune Lehmann*
Informatics and Mathematical Modeling, Technical University of Denmark, Building 321, DK-2800 Kgs. Lyngby, Denmark.

Andrew D. Jackson and Benny E. Lautrup
The Niels Bohr Institute, Blegdamsvej 17, DK-2100 København Ø, Denmark.
(Dated: January 27, 2007)

Condensing the work of any academic scientist into a one-dimensional measure of scientific quality is a difficult problem. Here, we employ Bayesian statistics to analyze several different measures of quality. Specifically, we determine each measure's ability to discriminate between scientific authors. Using scaling arguments, we demonstrate that the best of these measures require approximately 50 paper to draw conclusions regarding long term scientific performance with usefully small statistical uncertainties. Further, the approach described here permits the value-free (i.e., statistical) comparison of scientists working in distinct areas of science.

PACS numbers: 89.65.-s,89.75.Da

I. INTRODUCTION

It appears obvious that a fair and reliable quantification of the 'level of excellence' of individual scientists is a near-impossible task [1–5]. Most scientists would agree on two qualitative observations: *(i)* It is better to publish a large number of articles than a small number. *(ii)* For any given paper, its citation count—relative to citation habits in the field in which the paper is published—provides a measure of its quality. It seems reasonable to assume that the quality of a scientist is a function of his or her full citation record[1]. The question is whether this function can be determined and whether quantitatively reliable rankings of individual scientists can be constructed. A variety of 'best' measures based on citation data have been proposed in the literature and adopted in practice [6, 7]. The specific merits claimed for these various measures rely largely on intuitive arguments and value judgments that are not amenable to quantitative investigation. (Honest people can disagree, for example, on the relative merits of publishing a single paper with 1000 citations and publishing 10 papers with 100 citations each.) The absence of quantitative support for any given measure of quality based on citation data is of concern since such data is now routinely considered in matters of appointment and promotion which affect every working scientist.

Citation patterns became the target of scientific scrutiny in the 1960s as large citation databases became available through the work of Eugene Garfield [8] and other pioneers in the field of bibliometrics. A surprisingly, large body of work on the statistical analysis of citation data has been performed by physicists. Relevant papers in this tradition include the pioneering work of D. J. de Solla Price, e.g. [9], and, more recently, [6, 10–12]. In addition, physicists are a driving force in the emerging field of complex networks. Citation networks represent one popular network specimen in which papers correspond to nodes connected by references (out-links) and citations (in-links). Citation networks have frequently been used as an example of growing networks with preferential attachment [13]. For reviews on this extensive subject, see [14–16]. The aim of the present paper is to take such studies in a novel direction by addressing the question of which one-dimensional measure of citation data is best in a manner which is both quantitative and free of value judgments. Given the remarks above, the ability to answer this question depends on a careful definition of the word 'best'.

The primary purpose of analyzing and comparing the citation records of individual scientists is to discriminate between them, i.e., to assign some measure of quality and its associated uncertainty to each scientist considered. Whatever the intrinsic and value-based merits of the measure, m, assigned to every author, it will be of no practical value unless the corresponding uncertainty, δm is sufficiently small. From this point of view, the best choice of measure will be that which provides maximal discrimination between scientists and hence the smallest value of δm. We will demonstrate that the question of deciding which of several proposed measures is most discriminating, and therefore 'best', can be addressed quantitatively using standard statistical methods.

Although the approach is straightforward, it is useful first to describe it in general. We begin by binning all authors by some tentative measure, m, of the quality of their full citation record. The probability that an author will lie in bin α is denoted $p(\alpha)$. Similarly, we bin each paper according to the total number of citations[2]. The full citation record for an author is simply the set $\{n_i\}$, where n_i is the number of his/her paper in citation bin i. For each author bin, α, we then empirically construct the conditional probability distribution, $P(i|\alpha)$, that a single paper by an author in this bin will lie in citation bin i. These conditional probabilities are the central ingredient in our analysis. They can be used to calculate the probability, $P(\{n_i\}|\alpha)$, that any full citation record was actually drawn at random on the conditional distribution, $P(i|\alpha)$ appropriate for

*Electronic address: slj@imm.dtu.dk
[1] Citation data is, in fact, publicly available for all academic scientists.
[2] We use the Greek alphabet when binning with respect to to m and the Roman alphabet for binning citations.

a fixed author bin, α. Bayes' theorem allows us to invert this probability to yield

$$P(\alpha|\{n_i\}) \sim P(\{n_i\}|\alpha)\, p(\alpha)\ , \qquad (1)$$

where $P(\alpha|\{n_i\})$ is the probability that the citation record $\{n_i\}$ was drawn at random from author bin α. By considering the actual citation histories of authors in bin β, we can thus construct the probability $P(\alpha|\beta)$, that the citation record of an author initially assigned to bin β was drawn on the the distribution appropriate for bin α. In other words, we can determine the probability that an author assigned to bin β on the basis of the tentative quality measure should actually be placed in bin α. This allows us to determine both the accuracy of the initial author assignment its uncertainty in a purely statistical fashion.

While a good choice of measure will assign each author to the correct bin with high probability this will not always be the case. Consider extreme cases in where we elect to bin authors on the basis of measures unrelated to scientific quality, e.g., by hair/eye color or alphabetically. For such measures $P(i|\alpha)$ and $P(\{n_i\}|\alpha)$ will be independent of α, and $P(\alpha|\{n_i\})$ will become proportional to prior distribution $p(\alpha)$. As a consequence, the proposed measure will have no predictive power whatsoever. It is obvious, for example, that a citation record provides no information of its author's hair/eye color. The utility of a given measure (as indicated by the statistical accuracy with which a value can be assigned to any given author) will obviously be enhanced when the basic distributions $P(i|\alpha)$ depend strongly on α. These differences can be formalized using the standard Kullback-Leibler divergence. As we shall see, there are significant variations in the predictive power of various familiar measures of quality.

The organization of the paper is as follows. Section II is devoted to a description of the data used in the analysis, Section III introduces the various measures of quality that we will consider. In Sections IV and V, we provide a more detailed discussion of the Bayesian methods adopted for the analysis of these measures and a discussion of which of these measures is best in the sense described above of providing the maximum discriminatory power. This will allow us in Section VI to address to the question of how many papers are required in order to make reliable estimates of a given author's scientific quality; finally, Section A discusses the origin of asymmetries in some the measures. A discussion of the results and various conclusions will be presented in Section VII.

II. DATA

The analysis in this paper is based on data from the SPIRES[3] database of papers in high energy physics. Our data set consists of all citable papers written by academic scientists from the theory subfield, ultimo 2003. All citations to papers outside of SPIRES were removed. In the context of this paper, we define an academic scientist as someone who has published 25 papers or more. This definition is intended to include almost everyone with a permanent academic position and exclude those who leave academia early in their careers (and generally cease active journal publication) in the interests of maintaining the homogeneity of the data sample. For more see [17], Chapters 3 and 4. The resulting data set includes 6 737 authors and a total of 274 470 papers. The actual number of papers is smaller than this since each multiple author paper is counted once per co-author. The theory subfield is, however, that part of high energy physics where this effect is least pronounced. This is due to the relatively small number of co-authors (typically $1-3$) per theoretical paper. In the case of the theory subfield, this weighting of papers by the number of co-authors has been shown to have negligible effects [11].

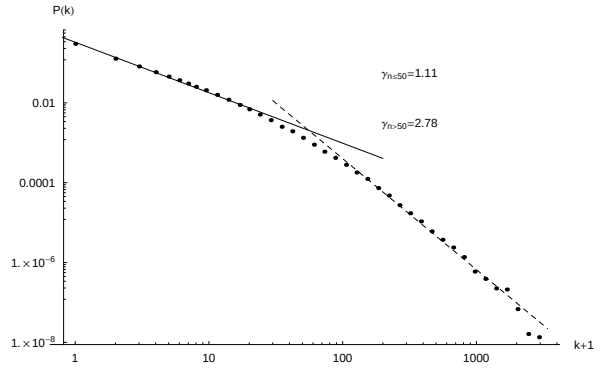


FIG. 1: Logarithmically binned histogram of the citations counts of all papers by authors with more than 25 publications in the theory subsection of SPIRES. The data is normalized and the axes are logarithmic.

The theory subsection of the SPIRES data has a power-law structure. Specifically the probability that a paper will recieve k citations is approximately proportional to $(k+1)^{-\gamma}$ with $\gamma = 1.11$ for $k \leq 50$ and $\gamma = 2.78$ for $k > 50$. The transition between these two power laws is found to be surprisingly sharp [11]. These features of the global distribution are also present in the conditional probabilities for subgroups of authors binned according to most measures of quality. In virtually all cases, these conditional probabilities can also be described accurately by separate power laws in each of two regions with a relatively sharp transition between the regions. As one might expect, authors with more citations are described by flatter distributions (i.e., smaller values of γ) and a somewhat higher transition point. Figure 1 displays the total distribution of citations as a binned and normalized

[3] SPIRES is an acronym for Stanford Physics Information REtrieval System. The database is open and can be found at http://www.slac.stanford.edu/spires/. Citations in SPIRES are gathered only from the papers in the database that have references entered electronically via eprints or journal articles, publications such as monographs or conference proceedings are treated inconsistently and therefore not included in this study.

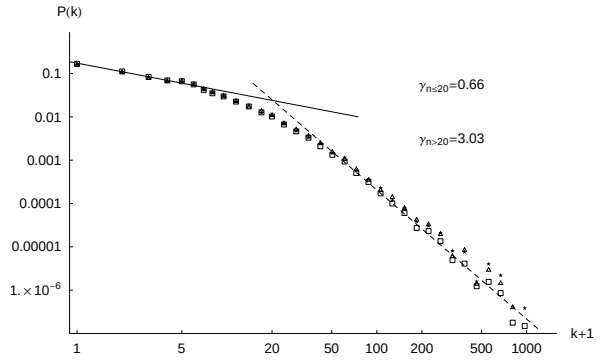


FIG. 2: Logarithmically binned histogram of the citations in bin 6 of the median measure. The △ points show the citation distribution of the first 25 papers by all authors. The points marked by ★ show the distribution of citations from the first 50 papers by authors who have written more than 50 papers. Finally, the □ data points show the distribution of all papers by all authors. The axes are logarithmic.

histogram[4].

Studies performed on the first 25, first 50 and all papers for a given value of m show the absence of temporal correlations. It is of interest to see this explicitly. Consider the following example. In Figure 2, we have plotted the distribution for bin 6 of the median measure[5]. There are 674 authors in this bin. Two thirds of these authors have written 50 papers or more. Only this subset is used when calculating the first 50 papers results. In this bin, the means for the total, first 25 and first 50 papers are 11.3, 12.8, and 12.9 citations per paper, respectively. The median of the distributions are 4, 6, and 6. The plot in Figure 2 confirms these observations. The remaining bins and the other measures yield similar results.

Note that Figure 2 confirms the general observations on the shapes of the conditional distributions made above. Figure 2 also shows two distinct power-laws. Both of the power-laws in this bin are flatter than the ones found in the total distribution and the transition point is lower than in the total distribution from Figure 1.

III. MEASURES OF SCIENTIFIC EXCELLENCE

Despite differing citation habits in different fields of science, most scientists agree that the number of citations of a given paper is the best objective measure of the quality of that paper. The belief underlying the use of citations as a measure of quality is that the number of citations to a paper provides an indication of how often the content of that paper has been used in the work of others[6]. Note, however, the obvious fact that citations can only be interpreted as a meaningful proxy of quality relative to the citation habits of one's peers or, put slightly differently, in the context of the citation habits of the field in which the paper is published. In [11], we have shown that the theory subsection of SPIRES is indeed a very homogeneous data set. In this sense, we will assume that the citation count of a paper is a proxy of the intrinsic quality of that paper.

The questions remain, however, of how to extract a measure of the quality of an individual scientist from his citation record and how fairly to project this record onto a scalar measure. This question is non-trivial because the probability, $p(k)$ of finding a scientific paper with k citations roughly follows an asymptotic power-law distribution, see Figs. 1 and 2. This fact was documented for the SPIRES data in Ref. [11] and holds true in many other scientific fields [9, 10, 16]. Thus, it is useful to consider some of the properties of the distribution of citations for all authors before discussing the various specific measures of quality to be considered here.

Empirical evidence indicates that most citation distributions are largely power-law distributed with $p(k) \sim k^{-\gamma}$. For small values of k, $\gamma \approx 1$; for larger values, $2 < \gamma < 3$. Although the average number of citations per paper is well-defined, the asymptotic power-law tails of these distributions cause their variance to be infinite[7]. When the variance is not defined (or very large), the mean values of a finite sample fluctuate significantly as a function of sample size. As a consequence, the average number of citations, $\langle k \rangle$, in the citation record of a given author (which is precisely a finite sample drawn from a power-law probability distribution) is a potentially unreliable measure of the quality of an author's citation record since the addition or removal of a single highly cited paper can materially alter an author's mean. Nevertheless, the mean of an author's citations is commonly used as an intensive scalar measure of author quality.

The reservations just expressed about the use of mean citations per paper apply with even greater force if one chooses to measure author quality by the number of citations of each author's single most highly cited paper, $k_{\max}$. Virtually all of the stabilizing statistical power of the full citation record has been discarded, and even greater fluctuations can be expected in this measure as the sample size changes. In spite of such statistical arguments, there are reasons for considering the maximum cited paper as a measure of quality. It is perfectly tenable to claim that the author of a single paper with

[4] Due to matters of visual presentation, the binning used in this and the following figure here is different from the binning used when constructing the $P(i|\alpha)$ used later in the paper. The correct binning is described in AppendixB

[5] Since this plot is constructed from authors assigned to bin 6, each paper is weighted by the number of its authors present in this bin. Weighing papers by the number of co-authors, however, does not significantly change the distribution of citations [11].

[6] We realize that there are a number of problems related to the use of citations as a proxy for quality. Papers may be cited or not for reasons other than their high quality. Geo- and/or socio-political circumstances can keep works of high quality out of the mainstream. Credit for an important idea can be attributed incorrectly. Papers can be cited for historical rather than scientific reasons. Indeed, the very question of whether authors actually read the papers they cite is not a simple one [18]. Nevertheless, we assume that correct citation usage dominates the statistics.

[7] Diverging higher moments of power-law distributions are discussed in the literature. E.g. [19].

1000 citations is of greater value to science than the author of 10 papers with 100 citations each (even though the latter is far less probable than the former). In this sense, the maximally cited paper might provide better discrimination between authors of 'high' and 'highest' quality, and this measure merits consideration.

Another simple and widely used measure of scientific excellence is the average number of papers published by an author per year. This would be a good measure if all papers were cited equally. As we have just indicated, scientific papers are emphatically not cited equally, and few scientists hold the view that all published papers are created equal in quality and importance. Indeed, roughly 50% of all papers in SPIRES are cited ≤ 2 times (including self-citation). This fact alone is sufficient to invalidate publication rate as a measure of scientific excellence. If all papers were of equal merit, citation analysis would provide a measure of industry rather than one of intrinsic quality.

In an attempt order to remedy this problem, *Thomson Scientific* (ISI) introduced the *Impact Factor*[8] which is designed to be a "measure of the frequency with which the 'average article' in a journal has been cited in a particular year or period"[9]. The Impact Factor can be used to weight individual papers. Unfortunately, citations to articles in a given journal also obey power-law distributions [12]. This has two consequences. First, the determination of the Impact Factor is subject to the large fluctuations which are characteristic of power-law distributions. Second, the tail of power-law distributions displaces the mean citation to higher values of k so that the majority of papers have citation counts that are much smaller than the mean. This fact is for example expressed in the large difference between mean and median citations per paper. For the total SPIRES data base, the median is 2 citations per paper; the mean is approximately 15. Indeed, only 22% of the papers in SPIRES have a number of citations in excess of the mean, cf. [11]. Thus, the dominant role played by a relatively small number of highly cited papers in determining the Impact Factor implies that it is subject to relatively large fluctuations and that it tends overestimate the level of scientific excellence of high impact journals. This fact was directly verified by Seglen [20], who showed explicitly that the citation rate for individual papers is uncorrelated to the impact factor of the journal in which it was published.

An alternate way to measure excellence is to categorize each author by the median number of citations of his papers, $k_{1/2}$. Clearly, the median is far less sensitive to statistical fluctuations since all papers play an equal role in determining its value. To demonstrate the robustness of the median, it is useful to note that the median of $\mathcal{N} = 2N+1$ random draws on *any* normalized probability distribution, $q(x)$, is normally distributed in the limit $\mathcal{N} \to \infty$. To this end we define the integral of $q(x)$ as

$$Q(x) = \int^x q(x')dx' \tag{2}$$

Evidently, $Q(x)$ grows monotonically from 0 to 1 independent of $q(x)$. The 'median' of this sample is defined as that value of x such that *(i)* one draw has the value x, *(ii)* N draws have a value less than or equal to x, and *(iii)* N draws have a value greater than or equal to x. The probability that the median is at x is now given as

$$P_{x_{1/2}}(x) = \frac{(2N+1)!}{1!N!N!} q(x)Q(x)^N[1-Q(x)]^N \,. \tag{3}$$

For large N, the maximum of $P_{x_{1/2}}(x)$ occurs at $x = x_{1/2}$ where $Q(x_{1/2}) = 1/2$. Expanding $P_{x_{1/2}}(x)$ about its maximum value, we see that

$$P_{x_{1/2}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x-x_{1/2})^2}{2\sigma^2}], \qquad \sigma^2 = \frac{1}{4q(x_{1/2})^2\mathcal{N}} \,. \tag{4}$$

A similar argument applies for every percentile. The statistical stability of percentiles suggests that they are well-suited for dealing with the power laws which characterize citation distributions.

Recently, Hirsch [6] proposed a different measure, h, intended to quantify scientific excellence. Hirsch's definition is as follows: "A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have fewer than h citations each"[6]. Unlike the mean and the median, which are intensive measures largely constant in time, h is an extensive measure which grows throughout a scientific career. Hirsch assumes that h grows approximately linearly with an author's professional age, defined as the time between the publication dates of the first and last paper. Unfortunately, this does not lead to an intensive measure. Consider, for example, the case of authors with large time gaps between publications, or the case of authors whose citation data are recorded in disjoint databases. A properly intensive measure can be obtained by dividing an author's h-index by the number of his/her total publications. We will consider both approaches below.

The h-index represents an attempt to strike a balance between productivity and quality and to escape the tyranny of power law distributions which place strong weight on a relatively small number of highly cited papers. The problem is that Hirsch assumes an equality between incommensurable quantities. An author's papers are listed in order of decreasing citations with paper i having $C(i)$ citations. Hirsch's measure is determined by the equality, $h = C(h)$, which posits an equality between two quantities with no evident logical connection. While it might be reasonable to assume that $h^\gamma \sim C(h)$, there is no reason to assume that γ and the constant of proportionality are both 1.

We will also include one intentionally nonsensical choice in the following analysis of the various proposed measures of author quality. Specifically, we will consider what happens when authors are binned alphabetically. In the absence of historical information, it is clear that an author's citation record

[8] For a full definition see http://scientific.thomson.com/knowtrend/essays/-journalcitationreports/impactfactor/.

[9] *Ibid.*

should provide us with no information regarding the author's name. Binning authors in alphabetic order should thus fail any statistical test of utility and will provide a useful calibration of the methods adopted. The measures of quality described in this section are the ones we will consider in the remainder of this paper.

IV. A BAYESIAN ANALYSIS OF CITATION DATA

The rationale behind all citation analyses lies in the fact that citation data is strongly correlated such that a 'good' scientist has a far higher probability of writing a good (i.e., highly cited) paper than a 'poor' scientist. Such correlations are clearly present in SPIRES [11, 21]. We thus categorize each author by some tentative quality index based on their total citation record. Once assigned, we can empirically construct the prior distribution, $p(\alpha)$, that an author is in author bin α and the probability $P(N|\alpha)$ that an author in bin α has a total of N publications. We also construct the conditional probability $P(i|\alpha)$ that a paper written by an author in bin α will lie in citation bin i. As we have seen earlier, studies performed on the first 25, first 50 and all papers of authors in a given bin reveal no signs of additional temporal correlations in the lifetime citation distributions of individual authors. In performing this construction, we have elected to bin authors in deciles. We bin papers into L bins according to the number of citations. The binning of papers is approximately logarithmic (see Appendix A). We have confirmed that the results stated below are largely independent of the bin-sizes chosen.

We now wish to calculate the probability, $P(\{n_i\}|\alpha)$, that an author in bin α will have the full (binned) citation record $\{n_i\}$. In order to perform this calculation, we assume that the various counts n_i are obtained from N *independent* random draws on the appropriate distribution, $P(i|\alpha)$. Thus,

$$P(\{n_i\}|\alpha) = P(N|\alpha)N!\prod_{i=1}^{L}\frac{P(i|\alpha)^{n_i}}{(n_i)!}\ . \qquad (5)$$

Although large scale temporal correlations are known to be absent, transient correlations are possible. For example, one particularly well-cited paper could lead to an increased probability of high citations for its immediate successor(s). It is difficult to demonstrate their presence or absence, but the results of following section will provide a posteriori evidence that such correlations, if present, are not important.

We can now invert the probability $P(\{n_j\}|\alpha)$ using Bayes' theorem to obtain

$$\begin{aligned} P(\alpha|\{n_i\}) &= \frac{P(\{n_i\}|\alpha)\,p(\alpha)}{P(\{n_i\})} \\ &= \frac{p(\alpha)P(N|\alpha)\prod_j P(j|\alpha)^{n_j}}{\sum_\beta p(\beta)P(N|\beta)\prod_k P(k|\beta)^{n_k}}\ , \end{aligned} \qquad (6)$$

where we have inserted Eq. (5) and used marginalization to obtain the normalization. The combinatoric factors cancel. The quantity $P(\alpha|\{n_i\})$, which represents the probability that an author with binned citation record $\{n_i\}$ is in author bin α. It can be used in two ways—each of which is interesting.

For any measure chosen Eq. (6) provides us with the probability that an author lies in author bin α. While the value of any measure (such as the mean number of citations per paper) can be calculated directly, the calculated values of $P(\alpha|\{n_i\})$ provide far more detailed and more reliable information using *all* statistical information contained in the data. The large fluctuations which can be encountered in identifying authors by their mean citation rate or by their maximally cited paper are reduced. Further, by providing us with values of $P(\alpha|\{n_i\})$ for all α, we obtain a statistically trustworthy gauge of whether the resulting uncertainties in α are sufficiently small for the measure under consideration to be a reliable indicator of author quality. In short, Eq. (6) provides us with a measure of an author's ranking independent of the total number papers currently published, and with information which allows us to assess the reliability of this determination. The accuracy of the resulting value of α increases dramatically with the total number of published papers. We will return to this point in Section V.

Fig. 3 shows the probabilities $P(\alpha|\{n_i\})$ that A will lie in each of the decile bins using the measures discussed in section II. These measures include: *(a)* the first initial of the author's name, *(b)* the average yearly output of papers, *(c)* Hirsch's h normalized by the author's professional age T, *(d)* the h-index normalized by the number of published papers, *(e)* the citation count of the single most cited paper, *(f)* the mean number of citations per paper, *(g)* the median number (50th percentile) of citations per paper, and *(h)* a 65th percentile measure. It is clear from the figure that there are significant differences, both in the accuracy of of the initial assignments and, more importantly, in the corresponding uncertainties. Large uncertainties are due to the fact that the conditional probabilities, $P(i|\alpha)$ are largely independent of α. Such independence is to be expected in the case of the alphabetic binning of authors, where the inability of the citation record to identify the first initial of author A's name is hardly surprising. The figure also suggests that the number of papers published per year is not reliable. Initial assignments of author A based on mean, median, 65th percentile, and maximum citations all appear to provide an accurate reflection of his full citation record with a satisfactorily small uncertainty. Hirsch's measures falls somewhere between the best and worst choice of measures.

Given the large variations in the accuracy and confidence of decile assignments as a function of the measure selected, it is of interest to investigate in greater detail the question of which of these measures is best. We address this question in the next section.

V. WEIGHING THE MEASURES

In order to obtain a more graphic representation of the quality of a given measure, we calculate the probability, $P(\beta|\alpha)$, that an author initially assigned to bin α is predicted to lie in bin β. In practice, we determine $P(\beta|\alpha)$ as the average of the probability distributions $P(\beta|\{n_i\})$ for each author in bin α. The results are shown 'stacked' in Fig. 4 for the various measures considered. Here, row α shows the (average) prob-

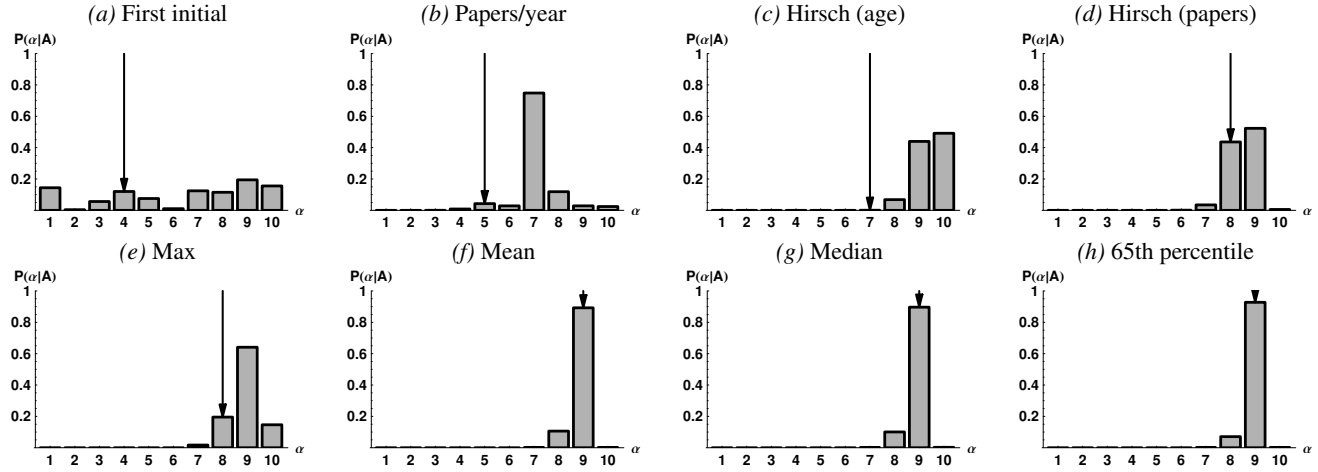


FIG. 3: A single author example. We analyze the citation record of author A with respect to the eight different measures defined in the text. Author A has written a total of 88 papers. The mean of this citation record is 26 citations per paper, the median is 13 citations, the h-index is 29, the maximally cited paper has 187 citations, and papers have been published at the average rate of 2.5 papers per year. The various panels show the probability that author A belongs to each of the ten deciles given on the corresponding measure; the vertical arrow displays the initial assignment. Panel *(a)* displays $P(\text{first initial}|A)$ *(b)* shows $P(\text{papers per year}|A)$, *(c)* shows $P(h/T|A)$, *(d)* shows $P(h/N|A)$, panel *(e)* shows $P(k_{\max}|A)$, panel *(f)* displays $P(\langle k\rangle|A)$, *(g)* shows $P(k_{1/2}|A)$, and finally *(h)* shows $P(k_{.65}|A)$.

abilities that an author initially assigned to bin α belongs in decile bin β. This probability is proportional to the area of the corresponding squares. Obviously, a perfect measure would place all of the weight in the diagonal entries of these plots. Weights should be centered about the diagonal for an accurate identification of author quality and the certainty of this identification grows as weight accumulates in the diagonal boxes. Note that an assignment of a decile based on Eq. (6) is likely to be more reliable than the value of the initial assignment since the former is based on all information contained in the citation record.

Figure 4 emphasizes that 'first initial' and 'publications per year' are not reliable measures. The h-index normalized by professional age performs poorly; when normalized by number of papers, the trend towards the diagonal is enhanced. We note the appearance of vertical bars in each figure in the top row. This feature is explained in Appendix A. All four measures in the bottom row perform fairly well. The initial assignment of the $k_{\max}$ measure always underestimates an author's correct bin. This is not an accident and merits comment. Specifically, if an author *has* produced a single paper with citations in excess of the values contained in bin α, the probability that he will lie in this bin, as calculated with Eq. (6), is strictly 0. Non-zero probabilities can be obtained only for bins including maximum citations greater than or equal to the maximum value already obtained by this author. (The fact that the probabilities for these bins shown in Fig. 4 are not strictly 0 is a consequence of the use of finite bin sizes.) Thus, binning authors on the basis of their maximally cited paper *necessarily* underestimates their quality. The mean, median and 65th percentile appear to be the most balanced measures with roughly equal predictive value.

It is clear from Eq. (6) that the ability of a given measure to discriminate is greatest when the differences between the conditional probability distributions, $P(i|\alpha)$, for different author bins are largest. These differences can quantified by measuring the 'distance' between two such conditional distributions with the aid of the Kullback-Leibler (KL) divergence (also know as the relative entropy). The KL divergence between two discrete probability distributions, p and p' is defined[10] as

$$\mathrm{KL}[p,p'] = \sum_i p_i \ln\left(\frac{p_i}{p'_i}\right). \tag{7}$$

The Kullback-Leibler divergence is positive and has desirable convexity properties. It is, however, not a metric due to the fact that $\mathrm{KL}[p',p] \neq \mathrm{KL}[p,p']$. While this asymmetry is of little concern when the differences between p and p' are small, some care is required when such differences are large. This can occur when the data set is so small that some citation bins are empty or when we bin authors by $k_{\max}$, in which case empty bins are inevitable as noted above. We consider the KL distance between adjacent distributions, Fig. 5 shows the distances $\mathrm{KL}[P(i|\alpha),P(i|\alpha+1)]$ for various measures. The probability $P(\beta=\alpha+1|\alpha)$ is exponentially sensitive to the KL divergence. Measures with large KL divergences between adjacent bins provide the most certain assignments of authors. The KL divergences for the measures not shown are significantly smaller than those displayed. The results of Fig. 5 provide quantitative support for the roughly equal performance of mean, median, and 65th percentile measures[11] seen in Figure 4. The h-index normalized by number of publications is

[10] The non-standard choice of the natural logarithm rather than the logarithm base two in the definition of the KL divergence, will be justified below.

[11] Figure 5 gives a misleading picture of the $k_{\max}$ measure, since the KL divergences $\mathrm{KL}[P(i|\alpha+1),P(i|\alpha)]$ are infinite as discussed above.

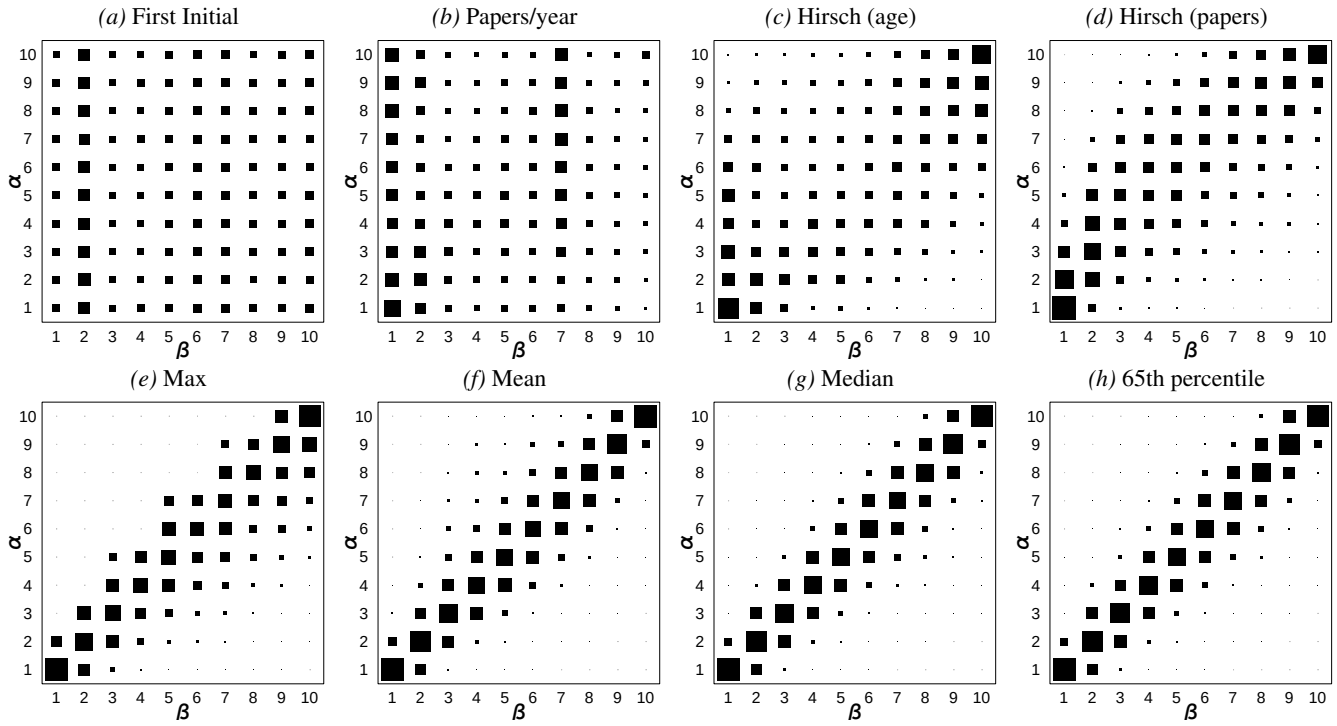


FIG. 4: Eight different measures. Each horizontal row shows the average probabilities (proportional to the areas of the squares) that authors initially assigned to decile bin α are predicted to belong in bin β. Panels as in Fig. 3.

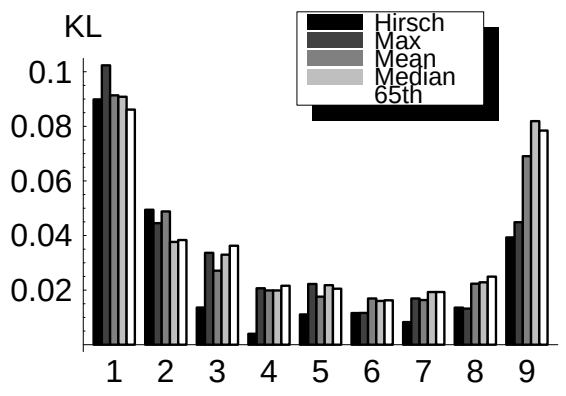


FIG. 5: The Kullback-Leibler divergences KL$[P(i|\alpha), P(i|\alpha+1)]$. Results are shown for the following distributions: *h*-index normalized by number of publications, maximum number of citations, mean, median, and 65th percentile.

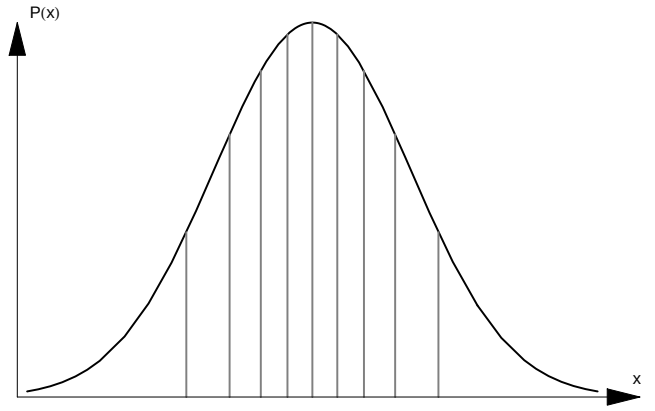


FIG. 6: Binning according to deciles. This plot displays a normal distribution (solid black line) as an example of a probability distribution peaked around a non-zero maximum. The grey vertical lines mark the boundaries of the 10 deciles.

dramatically smaller than the other measures shown except for the extreme deciles.

The reduced ability of all measures to discriminate in the middle deciles is immediately apparent from Fig. 5. This is a direct consequence any percentile binning given that the distribution of author quality has a maximum at some non-zero value, the bin size of a percentile distribution near the maximum will necessarily be small. The accuracy with which authors can be assigned to a given bin in the region around the maximum is reduced since one is attempting to distinguish between authors with very similar citation distributions. As a result, the statistical accuracy of percentile assignments is high at the extremes and relatively low in the middle of the distribution where we are attempting to make fine distinctions between scientists of similar ability. This effect is illustrated in Fig. 6.

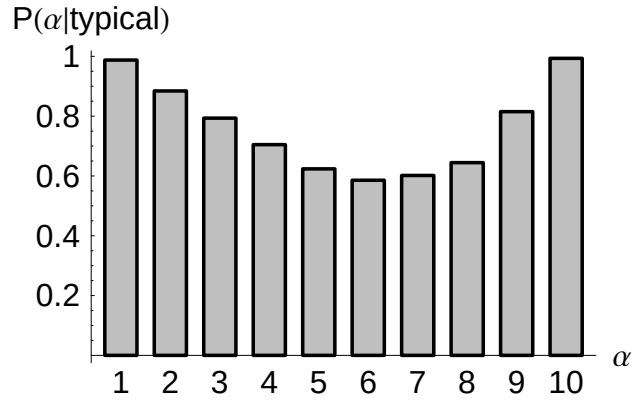


FIG. 7: The probability that a typical (i.e., most probable) author with 50 published papers will be assigned to the correct decile as a function of actual author decile. The median number of citations is used as a measure.

VI. SCALING

In this section, we consider the question of how many published papers are required in order to make a reliable prediction of the percentile ranking of a given author. (We consider results only using the 65th percentile measure.) If this number is sufficiently small, analysis along the lines presented here can provide a practical tool of potential value in predicting long-term scientific performance. In order to address this question, we will consider how $P(\alpha|\{n_i\})$ scales as a function of the total number of publications for an average author in each bin. Assume that an average author belonging to bin α draws N papers at random from the distribution of $P(i|\alpha)$. The most probable number of papers in each citation bin will thus be given as $n_i = NP(i|\alpha)$. Inserting this result into Eq. (6) and discarding all fixed factors, we find that

$$P(\alpha|\{n_i\}) \sim p(\alpha)\left(\prod_i P(i|\alpha)^{P(i|\alpha)}\right)^N . \qquad (8)$$

For the same citation record, $\{n_i\}$, a similar expression permits determination of the probability that this average author will be assigned to any bin, β. We see that

$$\lim_{N\to\infty}\frac{1}{N}\ln\left(\frac{P(\beta|\{n_i\})}{P(\alpha|\{n_i\})}\right) = -\mathrm{KL}\left[P(\bullet|\alpha),P(\bullet|\beta)\right] . \qquad (9)$$

This equation illustrates the utility of the KL divergence and explains the origin of its lack of symmetry. It is clear from Eqs. (8) and (9) that the probability of assigning this average author to the wrong bin will ultimately vanish exponentially with N. Given enough papers, the largest bin will ultimately dominate.

To obtain a quantitative sense of how many papers are required in practice, we pose the following question: What is the probability that a typical author from each author decile with $N = 50$ published papers will be assigned to the correct decile? The answer is plotted as a histogram in Fig. 7 using the 65th percentile citation rate as a measure (Similar results are obtained when using the mean or median citation rates). The figure indicates that $N = 50$ papers is more than sufficient to identify authors in the first and tenth deciles. In fact, approximately 25 and 20 papers respectively are sufficient to place authors in these deciles at the 90% confidence level. Fig. 7 also indicates that ≈ 50 published papers are sufficient to make meaningful assignments of authors to the second, third, and ninth deciles. All measures have difficulty in assigning authors to deciles $5-8$. As indicated by the small values of the KL divergence in these bins for all measures considered, the citation distributions of these authors are simply too similar to permit accurate discrimination (see arguments in the previous section). On the other hand, the probability that an author can be correctly assigned to one of these middle bins on the basis of 50 publication is high. This difficulty is due to the relatively small range of citations ranges which cover these bins: the 65th percentile-bins 5 though 8 contain authors with a 65th percentile between 5 and 13 citations (cf. the narrow ranges of the middle bins in the case of the mean, displayed in Table II).

VII. CONCLUSIONS

There are two distinct questions which must be addressed in any attempt to use citation data as an indication of author quality. The first is whether the measure chosen to characterize a given citation distribution or even the citation distribution itself reflects the qualities that we would like to probe. The second question is whether a given measure is capable of discriminating between authors in a statistically reliable way and, by extension, which of several measures is best. We have shown that the use of Bayesian statistics and the Kullback-Leibler divergence can answer this question in a value-neutral and statistically compelling manner. It is possible to draw reliable conclusions regarding an author's citation record on the basis of approximately 50 papers, and it is possible to assign meaningful statistical uncertainties to the results. The high level of discrimination obtained in the highest and lowest deciles provides indirect support for our assumption that an author's citation record is drawn at random from an appropriate conditional distribution and suggests that possible additional correlations in citation data are not important. Further, the difficulty in discriminating between authors in the middle deciles suggests that intrinsic author ability is peaked at some non-zero value.

The probabilistic methods adopted here permit meaningful comparison of scientists working in distinct areas with only minimal value judgments. It seems fair, for example, to declare equality between scientists in the same percentile of their peer groups. It is similarly possible to combine probabilities in order to assign a meaningful ranking to authors with publications in several disjoint areas. All that is required is knowledge of the conditional probabilities appropriate for each homogeneous subgroup.

We note, however, that the number of publications required to make meaningful author assignments is large enough to limit the utility of such analyses in the academic appointment

process. This raises the question of whether there are more efficient measures of an author's full citation record than those considered here. Our object has been to find that measure which is best able to assign the most similar authors together. Straightforward iterative schemes can be constructed to this end and are found to converge rapidly (i.e., exponentially) to an optimal binning of authors. (The result is optimal in the sense that it maximizes the sum of the KL divergences, $KL[P(\bullet|\alpha),P(\bullet|\beta)]$, over all α and β.) The results are only marginally better than those obtained here with the mean, median or 65th percentile measures.

Finally, it is also important to recognize that it takes time for a paper to accumulate its full complement of citations. While their are indications that an author's early and late publications are drawn (at random) on the same conditional distribution [11], many highly cited papers accumulate citations at a constant rate for many years after their publication. This effect, which has not been addressed in the present analysis, represents a serious limitation on the value of citation analyses for younger authors. The presence of this effect also poses the additional question of whether there are other kinds of statistical publication data that can deal with this problem. Co-author linkages may provide a powerful supplement or alternative to citation data. (Preliminary studies of the probability that authors in bins α and β will co-author a publication reveal a striking concentration along the diagonal $\alpha = \beta$.) Since each paper is created with its full set of co-authors, such information could be useful in evaluating younger authors. This work will be reported elsewhere.

APPENDIX A: VERTICAL STRIPES

The most striking feature of the calculated $P(\beta|\alpha)$ shown in Fig. 4 is presence of vertical 'stripes'. These stripes are most pronounced for the poorest measures and disappear as the reliability of the measure improves. Here, we offer a schematic but qualitatively reliable explanation of this phenomenon. To this end, imagine that each author's citation record is actually drawn at random on the true distributions $Q(i|A)$. For simplicity, assume that every author has precisely N publications, that each author in true class A has the same distribution of citations with $n_i^A = NQ(i|A)$, and that there are equal numbers of authors in each true author class. These authors are then distributed into author bins, α, according to some chosen quality measure. The methods of Sections IV and V can then be used to determine $P(i|\alpha)$, $P(\{n_i^{(A)}\}|\beta)$, $P(\beta|\{n_i^{(A)}\})$ and $P(\beta|\alpha)$. Given the form of the $n_i^{(A)}$ and assuming that N is large, we find that

$$P(\beta|\{n_i^{(A)}\}) \approx \exp\left(-N\mathrm{KL}[Q(\bullet|A),P(\bullet|\beta)]\right) \qquad \text{(A1)}$$

and

$$\tilde{P}(\beta|\alpha) \sim \sum_A P(A|\alpha)\exp\left(-N\mathrm{KL}[Q(\bullet|A),P(\bullet|\beta)]\right) , \qquad \text{(A2)}$$

where $P(A|\alpha)$ is the probability that the citation record of an author assigned to class α was actually drawn on $Q(i|A)$. The results of this approximate evaluation are shown in Fig. 8 and compared with the exact values of $P(\beta|\alpha)$ for the papers per year measure. The approximations do not affect the qualitative features of interest.

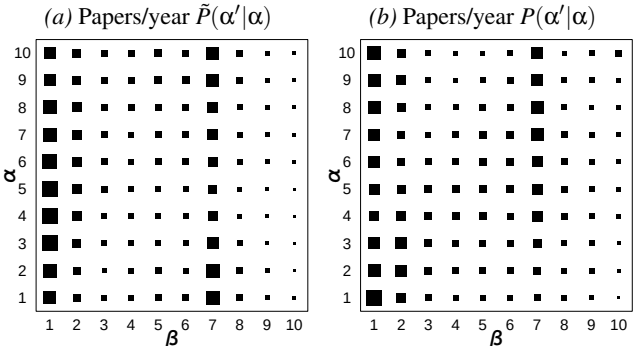


FIG. 8: A comparison of the approximate $\tilde{P}(\beta|\alpha)$ from Eq. (A2) and the exact $P(\beta|\alpha)$ for the papers published per year measure.

We now assume that the measure defining the author bins, α, provides a poor approximation to the true bins, A. In this case, authors will be roughly uniformly distributed, and the factor $P(A|\alpha)$ appearing in Eq. (A2) will not show large variations. Significant structure will arise from the exponential terms, where the presence of the factor N (assumed to be large), will amplify the differences in the KL divergences. The KL divergence will have a minimum value for some value of $A = A_0(\beta)$, and this single term will dominate the sum. Thus, $\tilde{P}(\beta|\alpha)$ reduces to

$$\tilde{P}(\beta|\alpha) \sim P(A_0|\alpha)\exp\left(-N\mathrm{KL}[Q(\bullet|A_0),P(\bullet|\beta)]\right) . \qquad \text{(A3)}$$

The vertical stripes prominent in Figs. 4(a) and (b) emerge as a consequence of the dominant β-dependent exponential factor. The present arguments also apply to the worst possible measure, i.e., a completely random assignment of authors to the bins α. In the limit of a large number of authors, N_{aut}, all $P(i|\beta)$ will be equal except for statistical fluctuations. The resulting KL divergences will respond linearly to these fluctuations.[12] These fluctuations will be amplified as before provided only that N_{aut} grows less rapidly than N^2. The argument here does *not* apply to good measures where there is significant structure in the term $P(A|\alpha)$. (For a perfect measure, $P(A|\alpha) = \delta_{A\alpha}$.) In the case of good measures, the expected dominance of diagonal terms (seen in the lower row of Fig. 4) remains unchallenged.

APPENDIX B: EXPLICIT DISTRIBUTIONS

For convenience we present all data to determine the probabilities $P(\alpha|\{n_i\})$ for authors who publish in the theory subsection of SPIRES. Data is presented only for case of the mean

[12] This is true because there will be no choice of A such that $Q(i|A) = P(i|\alpha)$.

$P(i\|\alpha)$		$P(N\|\alpha)$	
Bin number	Citation range	Bin Number	Total paper range
$i=1$	$k=1$	$m=1$	$N=25$
$i=2$	$k=2$	$m=2$	$N=26$
$i=3$	$2<k\leq 4$	$m=3$	$26<N\leq 28$
$i=4$	$4<k\leq 8$	$m=4$	$28<N\leq 32$
$i=5$	$8<k\leq 16$	$m=5$	$32<N\leq 40$
$i=6$	$16<k\leq 32$	$m=6$	$40<N\leq 56$
$i=7$	$32<k\leq 64$	$m=7$	$56<N\leq 88$
$i=8$	$64<k\leq 128$	$m=8$	$88<N\leq 152$
$i=9$	$128<k\leq 256$	$m=9$	$152<N\leq N_{\max}$
$i=10$	$256<k\leq 512$		
$i=11$	$512<k\leq k_{\max}$		

TABLE I: The binning of citations and total number of papers. The first and second column show the bin number and bin ranges for the citation bins used to determine the conditional citation probabilities $P(i|\alpha)$ for each α, shown in Table III. The third and fourth column display the bin number and total number of paper ranges used in the creation of the conditional probabilities $P(m|\alpha)$ for each α, displayed in Table IV.

α	$\langle k\rangle$-range	# authors	$p(\alpha)$	$\bar{n}(\alpha)$
1	0 – 1.69	673	0.1	37.0
2	1.69 – 3.08	673	0.1	41.8
3	3.08 – 4.88	675	0.1	44.0
4	4.88 – 6.94	673	0.1	46.8
5	6.94 – 9.40	674	0.1	52.2
6	9.40 – 12.56	674	0.1	54.3
7	12.56– 16.63	673	0.1	59.5
8	16.63– 22.19	674	0.1	59.0
9	22.19– 33.99	674	0.1	65.4
10	33.99–285.88	674	0.1	72.2

TABLE II: The author bins. This table shows the mean numbers of citations that define the limits of the 10 author bins.

[13] This fact is known as *Lotka's Law* [22].

number of citations. All citations are binned logarithmically according to the citation bins listed in column one and two of Table I. The author bins are determined on the basis of deciles of the total distribution of mean citations, $p(\langle k\rangle)$. Table II shows the relevant quantities for these bins. Given the definitions of both the author- and citation bins, we can determine the conditional citation distributions $P(i|\alpha)$ empirically. These are given in Table III.

We also need the probabilities $P(N|\alpha)$ describing that an author in bin α has N publications. Because of the low number of authors in each bin, we need to bin the total number of publications when calculating this probability; we use the letter m to enumerate the N-bins. Because $P(N|\alpha)$ is described by a power-law distribution[13] and since we only consider authors with more than 25 publications, we choose to bin N logarithmically as displayed in the third and fourth column of Table I. The conditional probabilities, $P(m|\alpha)$ are displayed in Table IV.

[1] D. Adam. The counting house. *Nature*, 415:726, 2002.

[2] R. P. Dellavalle, E. J. Hester, L. F. Heilig, A. L. Drake, J. W. Kuntzman, M. Graber, and L. M. Schilling. Going, going, gone: Lost internet references. *Science*, 302:787, 2003.

[3] G. Franck. Scientific communication – A vanity fair. *Science*, 286:53, 1999.

[4] J. Adams and Z. Griliches. Measuring science: An exploration. *Proceedings of the National Academy of Sciences, USA*, 93:12664, 1996.

[5] S. Lehmann, A. D. Jackson, and B. E. Lautrup. Measures for measures. *Nature*, 444:1003, 2006.

[6] J. E. Hirsch. An index to quantify an individual's scientific output. *Proceedings of the National Academy of the Sciences*, 102:16569, 2005.

[7] A. F. J. van Raan. Statistical properties of bibliometric indicators: Research group indicator distributions and correlations. *Journal of the American Society for Information Science*, 57:408, 2005.

[8] E. Garfield. *Essays of an Information Scientist*, volume 1-15. ISI Press, 1977-1993.

[9] D. de Solla Price. Networks of scientific papers. *Science*, 149:510, 1965.

[10] S. Redner. How popular is your paper? An emperical study of the citation distribution. *European Physics Journal B*, 4:131, 1998.

[11] S. Lehmann, B. E. Lautrup, and A. D. Jackson. Citation networks in high energy physics. *Physical Review E*, 68, 2003.

[12] S. Redner. Citation statistics from 110 years of physical review. *Physics Today*, 58:49, 2005.

[13] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[14] R. Albert and A.-L. Barabási. Statistical mechanics of complex

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$	$i=10$	$i=11$
$\alpha=1$	0.612	0.182	0.127	0.057	0.019	0.002	0.000	0.000	0.000	0.000	0.000
$\alpha=2$	0.433	0.188	0.181	0.122	0.055	0.016	0.004	0.000	0.000	0.000	0.000
$\alpha=3$	0.327	0.165	0.188	0.167	0.103	0.038	0.010	0.002	0.000	0.000	0.000
$\alpha=4$	0.263	0.143	0.178	0.184	0.140	0.067	0.019	0.005	0.001	0.000	0.000
$\alpha=5$	0.217	0.127	0.163	0.183	0.165	0.096	0.036	0.009	0.002	0.000	0.000
$\alpha=6$	0.177	0.113	0.150	0.181	0.173	0.126	0.058	0.017	0.004	0.001	0.000
$\alpha=7$	0.143	0.098	0.135	0.170	0.183	0.149	0.086	0.028	0.007	0.002	0.000
$\alpha=8$	0.118	0.080	0.121	0.155	0.182	0.169	0.110	0.048	0.012	0.003	0.000
$\alpha=9$	0.094	0.066	0.099	0.141	0.175	0.178	0.139	0.075	0.025	0.007	0.001
$\alpha=10$	0.068	0.045	0.071	0.107	0.145	0.171	0.166	0.121	0.067	0.027	0.012

TABLE III: The distributions $P(i|\alpha)$. This table displays the conditional probabilities that an author writes a paper in paper-bin i given that his author-bin is α.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$	$m=9$
$\alpha=1$	0.083	0.071	0.134	0.226	0.224	0.172	0.082	0.006	0.001
$\alpha=2$	0.058	0.049	0.103	0.187	0.236	0.217	0.122	0.025	0.003
$\alpha=3$	0.068	0.050	0.095	0.133	0.231	0.240	0.136	0.041	0.004
$\alpha=4$	0.043	0.049	0.095	0.141	0.198	0.247	0.162	0.061	0.004
$\alpha=5$	0.031	0.059	0.067	0.108	0.181	0.246	0.200	0.091	0.016
$\alpha=6$	0.031	0.039	0.068	0.126	0.162	0.245	0.215	0.099	0.015
$\alpha=7$	0.034	0.022	0.058	0.114	0.152	0.242	0.215	0.128	0.034
$\alpha=8$	0.028	0.024	0.049	0.096	0.178	0.243	0.248	0.101	0.033
$\alpha=9$	0.030	0.033	0.037	0.074	0.148	0.228	0.245	0.160	0.045
$\alpha=10$	0.027	0.028	0.043	0.077	0.131	0.212	0.199	0.223	0.061

TABLE IV: The conditional probabilities $P(m|\alpha)$. This table contains the conditional probabilities that an author has a total number of publications in publication-bin m given that his author-bin is α .

networks. *Reviews of modern physics*, 74:47, 2002.

[15] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51:1079, 2002.

[16] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[17] S. Lehmann. Spires on the building of science. Master's thesis, The Niels Bohr Institute, 2003. May be downloaded from www.imm.dtu.dk/∼slj/.

[18] M. V. Simkin and V. P. Roychowdhury. Read before you cite! *Complex Systems*, 14:269, 2003.

[19] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323, 2005.

[20] P. O. Seglen. Casual relationship between article citedness and journal impact. *Journal of the American Society for Information Science*, 45:1, 1994.

[21] S. Lehmann, A. D. Jackson, and B. E. Lautrup. Life, death, and preferential attachment. *Europhysics Letters*, 69:298, 2005.

[22] A. J. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16:317, 1926.

CHAPTER 6

Community Structure

THE measures of quality from the previous chapter create a ranking of the author publication records $\{k_i\}$. By forming communities of nodes in the author-network based on this ranking, Lehmann *et al*. [56, 57] argue that a good measure of quality is one that creates groups, where all nodes have similar citation records and where the groups themselves are as heterogeneous as possible. A bad measure corresponds to grouping authors at random.

In this chapter, we will continue to concentrate on communities, but consider only networks that are much simpler than the author-network in SPIRES. We shall return to networks that are simple in the sense that they can be described completely with adjacency matrices that are symmetric and contain only 0s and 1s.

6.1 Deterministic Modularity Optimization

The paper *Deterministic Modularity Optimization* [52] regards maximization of the modularity Q defined in equation (2.27). Lehmann and Hansen propose a

novel scheme for the optimization of Q based on deterministic mean field methods. Further, these authors propose a simple class of random networks with adjustable community structure. Given a set of parameters that characterize the network model, they demonstrate how to calculate the modularity of this particular network model, analytically.

The simple model is used as a testing ground for the mean field optimization and the mean field scheme is shown to find higher values of Q for all tested parameter settings than any previously used deterministic optimization methods.

This paper has recently been submitted to the journal *Physical Review E* and is currently available from the online preprint server http://arXiv.org/ under the label *physics/0701348*.

Deterministic Modularity Optimization

Sune Lehmann* and Lars Kai Hansen
Informatics and Mathematical Modeling, Technical University of Denmark, Building 321, DK-2800 Kgs. Lyngby, Denmark.
(Dated: January 31, 2007)

We study community structure of networks. We have developed a scheme for maximizing the modularity Q [1] based on mean field methods. Further, we have defined a simple family of random networks with community structure; we understand the behavior of these networks analytically. Using these networks, we show how the mean field methods display better performance than previously known deterministic methods for optimization of Q.

PACS numbers:

I. INTRODUCTION

A theoretical foundation for understanding complex networks has developed rapidly over the course of the past few years [2–4]. More recently, the subject of detecting network *communities* has gained an large amount of attention, for reviews see Refs [5, 6]. Community structure describes the property of many networks that nodes divide into modules with dense connections between the members of each module and sparser connections between modules.

In spite of a tremendous research effort, the mathematical tools developed to describe the structure of large complex networks are continuously being refined and redefined. Essential features related to network structure and topology are not necessarily captured by traditional global features such as the average degree, degree distribution, average path length, clustering coefficient, etc. In order to understand complex networks, we need to develop new measures that capture these structural properties. Understanding community structures is an important step towards developing a range of tools that can provide a deeper and more systematic understanding of complex networks. One important reason is that modules in networks can show quite heterogenic behavior [7], that is, the link structure of modules can vary significantly from module to module. For such heterogenic systems, global measures can be directly misleading. Also, in practical applications of network theory, knowledge of the community structure of a given network is important. Access to the modular structure of the internet could help search engines supply more relevant responses to queries on terms that belong to several distinct communities[1]. In biological networks, modules can correspond to functional units of some biological system [8].

II. THE MODULARITY

This section is devoted to an analysis of the modularity Q. Identifying communities in a graph has a long history in mathematics and computer science [5, 9]. One obvious way to partition a graph into C communities is distribute nodes into the communities, such that the number of links connecting the different modules of the network is minimized. The minimal number of connecting links is called the *cut size* R of the network.

Consider an unweighted and undirected graph with n nodes and m links. This network can be represented by an adjacency matrix $\mathbf{A}$ with elements

$$A_{ij} = \begin{cases} 1, & \text{if there is a link joining nodes } i \text{ and } j; \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

This matrix is symmetric with $2m$ entries. The degree k_i of node i is given by $k_i = \sum_j A_{ij}$. Let us express the cut-size in terms of $\mathbf{A}$; we find that

$$R = \frac{1}{2}\sum_{i,j} A_{ij}[1-\delta(c_i,c_j)], \tag{2}$$

where c_i is the community to which node i belongs and $\delta(\alpha,\beta) = 1$ if $\alpha = \beta$ and $\delta(\alpha,\beta) = 0$ if $\alpha \neq \beta$. Minimizing R is an integer programming problem that can be solved exactly in polynomial time [10]. The leading order of the polynomial, however, is n^{C^2} which very expensive for even very small networks. Due to this fact, most graph partitioning has been based on spectral methods (more below).

Newman has argued [5, 7, 11] that R is not the right quantity to minimize in the context of complex networks. There are several reasons for this: First of all, the notion of cut-size does not capture the essence of our 'definition' of network as a tendency for nodes to divide into modules with dense connections between the members of module and sparser connections between modules. According to Newman, a good division is not necessarily one, in which there are few edges between the modules, it is one where there are fewer edges than expected. There are other problems with R: If we set the community sizes free, minimizing R will tend to favor small communities, thus the use of R forces us to decide on and set the sizes of the communities in advance.

As a solution to these problems, Girvan and Newman propose the modularity Q of a network [1], defined as

$$Q = \frac{1}{2m}\sum_{ij}[A_{ij} - P_{ij}]\delta(c_i,c_j). \tag{3}$$

*Electronic address: slj@imm.dtu.dk

[1] Some search engines have begun implementing related ideas, see for example *Clusty, the Clustering Engine* (http://clusty.com/). There is, however, still considerable room for improvement.

The P_{ij}, here, are a null model, designed to encapsulate the 'more edges than expected' part of the intuitive network definition. It denotes the probability that a link exists between node i and j. Thus, if we know nothing about the graph, an obvious choice would be to set $P_{ij} = p$, where p is some constant probability. However, we know that the degree distributions of real networks are often far from random, therefore the choice of $P_{ij} \sim k_i k_j$ is sensible; this model implies that the probability of a link existing between two nodes is proportional to the degree of the two nodes in question. We will make exclusive use of this null model in the following; the properly normalized version is $P_{ij} = (k_i k_j)/(2m)$. It is axiomatically demanded that that $Q = 0$ when all nodes are placed in one single community. This constrains the P_{ij} such that

$$\sum_{ij} P_{ij} = 2m, \tag{4}$$

we also note that $\mathbf{P} = (\mathbf{P})^T$, which follows from the symmetry of $\mathbf{A}$.

Comparing Eqs. (2) and (3), we notice that there are two differences between Q and R. The first is that Q implies that we *maximize* the number of intra-community links instead of minimizing the the number of inter-community links as is the case for R—this is the difference between multiplying by $\delta(c_i, c_j)$ and $[1 - \delta(c_i, c_j)]$. The second difference lies in the the introduction of the P_{ij} in Equation (3). The subtraction of P_{ij} serves to incorporate information about the inter-community links into the quantity we are optimizing.

Use of modularity to identify network communities is not, however, completely unproblematic. Criticism has been raised by Fortunato and Barthélemy [12] who point out that the Q measure has a resolution limit. This stems from the fact that the null model $P_{ij} \sim k_i k_j$ can be misleading. In a large network, the expected number of links between two small modules is small and thus, a single link between two such modules is enough to join them into a single community. A variation of the same criticism has been raised by Rosvall and Bergstrom [13]. These authors point out that the normalization of P_{ij} by the total number of links m has the effect that if one adds a distinct (not connected to the remaining network) module to the network being analyzed and partition the whole network again allowing for an additional module, the division of the original modules can shift substantially due to the increase of m.

In spite of these problems, the modularity is a highly interesting method for detecting communities in complex networks when we assume that the communities are similar in size. What makes the modularity particularly interesting compared to other clustering methods is its ability to inform us of the optimal number of communities for a given network[2].

[2] This ability to estimate the number of communities, however, stems from the introduction of the P_{ij} term in the Eq. (3) and is therefore directly linked to the conceptual problems with Q mentioned in the previous paragraph.

III. SPECTRAL OPTIMIZATION OF MODULARITY

The question of finding the optimal Q is a discrete optimization problem. We can estimate the size of the space we must search to find the maximum. The number of ways to divide n vertices into C non-empty sets (communities) is given by the Stirling number of the second kind $S_n^{(C)}$ [14]. Since we do not know the number of communities that will maximize Q before we begin dividing the network, we need to examine a total of $\sum_{C=2}^{n} S_n^{(C)}$ community divisions [15]. Even for small networks, this is an enormous space, which renders exhaustive search out of the question.

Motivated by the success of spectral methods in graph partitioning, Newman suggests a spectral optimization of Q [11]. We define a matrix, called the modularity matrix $\mathbf{B} = \mathbf{A} - \mathbf{P}$ and an $(n \times C)$ *community matrix* $\mathbf{S}$. Each column of $\mathbf{S}$ corresponds to a community of the graph and each row corresponds to a node, such that the elements

$$S_{ic} = \begin{cases} 1, & \text{if node } i \text{ belongs to community } c; \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Since each node can only belong to one community, the columns of $\mathbf{S}$ are orthogonal and $\mathrm{Tr}(\mathbf{S}^T\mathbf{S}) = n$. The δ-symbol in Equation (3) can be expressed as

$$\delta(c_i, c_j) = \sum_{k=1}^{C} S_{ik} S_{jk}, \tag{6}$$

which allows us to express the modularity compactly as

$$Q = \frac{1}{2m} \sum_{i,j=1}^{n} \sum_{k=1}^{C} B_{ij} S_{ik} S_{jk} = \frac{\mathrm{Tr}(\mathbf{S}^T\mathbf{B}\mathbf{S})}{2m}. \tag{7}$$

This is the quantity that we wish to maximize.

The next step is the 'spectral relaxation', where we relax the discreteness constraints on $\mathbf{S}$, allowing elements of this matrix to possess real values. We do, however, constrain the length of the column vectors by $\mathbf{S}^T\mathbf{S} = \mathbf{M}$, where $\mathbf{M}$ is a $C \times C$ matrix with the number of nodes in each community $n_1, n_2, \ldots, n_C$ along the diagonal. In order to determine the maximum, we take

$$\frac{\partial}{\partial \mathbf{S}} \left(\frac{1}{2m} \mathrm{Tr}[\mathbf{S}^T\mathbf{B}\mathbf{S}] + \mathrm{Tr}[(\mathbf{S}^T\mathbf{S} - \mathbf{M})\tilde{\Lambda}] \right) = 0, \tag{8}$$

where $\tilde{\Lambda}$ is a $C \times C$ diagonal matrix of Lagrange multipliers. The maximum is given by

$$\mathbf{B}\mathbf{S} = \mathbf{S}\Lambda, \tag{9}$$

where $\Lambda = -2m\tilde{\Lambda}$ for cosmetical reasons. Eq. (9) is a standard matrix eigenvalue problem. Optimizing in the relaxed representation, we substitute this solution into Eq. (7), and see that in order to maximize Q, we must choose the C largest eigenvalues of $\mathbf{B}$ and their corresponding eigenvectors. Since all rows and columns of $\mathbf{B}$ sum to zero by definition, the vector $(1, 1, \ldots, 1)^T$ is always an eigenvector of $\mathbf{B}$ with the eigenvalue 0. In general the modularity matrix can have both positive and negative eigenvalues. It is clear from Eq. (7) that the

eigenvectors corresponding to negative eigenvalues can never yield a positive contribution to the modularity. Thus, the number of positive eigenvalues presents an upper bound on the number of possible communities.

However, we need to convert our problem back to a discrete one. This is a non-trivial task. There is no standard way to go from the n continuous entries in each of the C largest eigenvectors of the modularity matrix and back to discrete $0,1$ values of the community matrix $\mathbf{S}$. One simple way of circumventing this problem is to use repeated bisection of the network. This is the procedure that Newman [11] recommends. In Newman's scheme, the only eigenvector utilized is the eigenvector corresponding to the largest eigenvalue $b_{\max}$ of $\mathbf{B}$ (with highest contribution to Q). The $0,1$ vector most parallel to this continuous eigenvector, is one where the positive elements of the eigenvector are set to one and the negative elements zero. This is the first column of the community matrix $\mathbf{S}$. The second column must contain the remaining elements.

We can increase the modularity iteratively by bisecting the network into smaller and smaller pieces. However, this repeated bisection of the network is problematic. There is no guarantee that that the best division into three groups can be arrived at by finding by first determine the best division into two and then dividing one of those two again. It is straight forward to construct examples where a sub-optimal division into communities is obtained when using bisection [7, 16].

Spectral optimization is not perfect—especially when only the eigenvector corresponding to $b_{\max}$ is employed[3]. Therefore, Newman suggests that it should only be used as a starting point. In order to improve the modularity, Newman has devised an algorithm inspired by the classical Kernighan-Lin (KL) scheme [17]. The procedure is as follows: After each bisection of the network we go through the nodes and find the one that yields the highest increase in the modularity of the entire network (or smallest decrease if no increase is possible) if moved to the other module. This node is now moved to the other module and becomes inactive. The next step is to go through the remaining $n-1$ nodes and perform the same action. We continue like this until all nodes have been moved. Finally, we go through all the intermediate states and pick the one with the highest value of Q. This is the new starting division. We proceed iteratively from this configuration until no further improvement can be found. Let us call this optimization the 'KLN-algorithm'.

In the spectral optimization, the computational bottleneck is the calculation of the leading eigenvector(s) of $\mathbf{B}$, which is non-sparse. Naively, we would expect this to scale like $O(n^3)$. However, $\mathbf{B}$'s structure allows for a faster calculation. We can write the product of $\mathbf{B}$ and a vector $\mathbf{v}$ [11] as

$$\mathbf{B}\mathbf{v} = \mathbf{A}\mathbf{v} - \frac{\mathbf{k}(\mathbf{k}^T\mathbf{v})}{2m}. \tag{10}$$

This way we have a divided the multiplication into (i) sparse matrix product with the adjacency matrix that takes $O(m+n)$, and (ii) the inner product $\mathbf{k}^T\mathbf{v}$ that takes $O(n)$. Thus the entire product $\mathbf{B}\mathbf{v}$ scales like $O(m+n)$. The total running for a bisection determining the eigenvector(s) is therefore $O((m+n)n)$ rather than the naive guess of $O(n^3)$. Using Eq. (10) during the KLN-algorithm reduces the cost of this step to $O((m+n)n)$ [11].

IV. MEAN FIELD OPTIMIZATION

Simulated annealing was proposed by Kirkpatrick *et al*. [18] who noted the conceptual similarity between global optimization and finding the ground state of a physical system. Formally, simulated annealing maps the global optimization problem onto a physical system by identifying the cost function with the energy function and by considering this system to be in equilibrium with a heat bath of a given temperature T. By annealing, i.e., slowly lowering the temperature of the heat bath, the probability of the ground state of the physical system grows towards unity. This is contingent on whether or not the temperature can be decreased slowly enough such that the system stays in equilibrium, i.e., that the probability is Gibbsian

$$P(\mathbf{S}|T) = \frac{1}{Z}\exp\left(-\frac{1}{T}Q(\mathbf{S})\right) = \frac{1}{Z}\exp\left(-\frac{\mathrm{Tr}(\mathbf{S}^T\mathbf{B}\mathbf{S})}{2m}\right). \tag{11}$$

Here, Z is a constant ensuring proper normalization. Kirkpatrick et al. realized the annealing process by Monte Carlo sampling. The representation of the constrained modularity optimization problem is equivalent to a C-state Potts model. Gibbs sampling for the Potts model with the modularity Q as energy function has been investigated by Reichardt and Bornholdt, see e.g., [16].

Mean field annealing is a deterministic alternative to Monte Carlo sampling for combinatorial optimization and has been pioneered by Peterson et al. [19, 20]. Mean field annealing avoids extensive stochastic simulation and equilibration, which makes the method particularly well suited for optimization. There is a close connection between Gibbs sampling and MF annealing. In Gibbs sampling, every variable is updated by random draw of a Potts state with a conditional distribution,

$$P(S_{i1},...,S_{iC}|\mathbf{S}_{\{-i\}},T) = \frac{P(\mathbf{S}|T)}{\sum_{S_{i1},...,S_{iC}} P(\mathbf{S}|T)}, \tag{12}$$

where the sum runs over the C values of the i'th Potts variable and $\mathbf{S}_{\{-i\}}$ denotes the set of Potts variables excluding the i'th node. As noted by [16], Eq. (12) is local in the sense that the part of the energy function containing variables not connected with the i'th cancels out in the fraction. The mean field approximation is obtained by computing the conditional mean

[3] Newman has proposed a scheme that utilizes two eigenvectors of the modularity matrix corresponding to the two highest eigenvalues [7] that—according to our experiments—performs slightly better than the single eigenvector method described above. However, after the application of the KLN-algorithm described in this section, we found no difference in the results found by using one or two eigenvectors.

of the set of variables coding for the i'th Potts variable using Eq. (12) and approximating the Potts variables in the conditional probability by their means [20]. This leads to a simple self-consistent set of non-linear equations for the means,

$$\mu_{ik} = \frac{\exp(\phi_{ik}/T)}{\sum_{k'=1}^{C}\exp(\phi_{ik'}/T)}, \quad \phi_{ik} = \sum_j B_{ij}\mu_{jk}. \tag{13}$$

For symmetric connectivity matrices with $\sum_j B_{ij} = 0$, the set of mean field equations has the unique high-temperature solution $\mu_{ik} = 1/C$. This solution becomes unstable at the mean field critical temperature, $T_c = b_{\max}/C$, determined by the maximal eigenvalue $b_{\max}$ of $\mathbf{B}$.

This mean field algorithm is fast. Each synchronous iteration (see Section VI for details on implementation) requires a multiplication of $\mathbf{B}$ by the mean vector μ. As we have seen, this operation can be performed in $O(m+n)$ time using the trick in Eq. (10). In these experiments, we have used a fixed number of iterations of the order of $O(n)$, which gives us a total of $O((m+n)n)$ similar to the case of by spectral optimization. (A forthcoming paper discusses the relationship between Gibbs sampling, mean field methods, and computational complexity.)

V. A SIMPLE NETWORK

We will perform our numerical experiments on a simple model of networks with communities. This model network consists of C communities with n_c nodes in each, the total network has $n = n_c C$ nodes. Without loss of generality, we can arrange our nodes according to their community; a sketch of this type of network is displayed in Figure 1. Communities

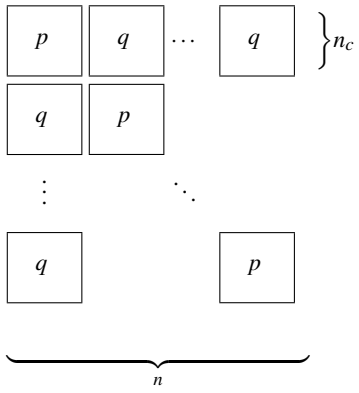


FIG. 1: A sketch of the simple network model. The figure displays the structure of the adjacency matrix with nodes arranged according to community. Inside each community (the blocks) along the diagonal, the probability of a link between two nodes is p and between communities, the probability of a link is q.

are defined as standard random networks, where the probability of a link between two nodes is given by p, with $0 < p \leq 1$. Between the communities the probability of a link between is given by $0 \leq q < p$. The networks are unweighted and undirected.

Let us calculate Q for this network in the case where $p = 1$ and $q = 0$. In this case, we can calculate everything exactly. First, we note that all nodes have the same number of links, and that the degree of node i, $k_i = n_c - 1$ (since a node does not link to itself). Thus the total number of links m_c in each sub-network is

$$m_c = \frac{1}{2}n_c(n_c-1), \tag{14}$$

and since our network consists of C identical communities the total number of links is $m = Cm_c$. We can now write down the contribution Q_c from each sub-network to the total modularity

$$\begin{aligned} Q_c &= \frac{1}{2m}\sum_{ij}(A_{ij}-P_{ij})\delta(c,c) &(15)\\ &= \frac{1}{2m}\left[n_c(n_c-1) - n_c^2\frac{(n_c-1)^2}{2m}\right]. &(16)\end{aligned}$$

If we insert m and use that $Q = CQ_c$, we find

$$Q = CQ_c = 1 - \frac{1}{C}. \tag{17}$$

We see explicitly that when $C \to \infty$ the modularity approaches unity.

Now, let us examine at the general case. Since our network is connected at random, we cannot calculate the number of links per node exactly, but we know that the network is well-behaved (Poisson link distribution), thus we can calculate the *average* number of links per node. We see that

$$k = (n_c-1)p + n_c(C-1)q, \tag{18}$$

which is equal to the number of expected intra-community links plus the number of expected number of inter-community links. The number of links in the entire network is therefore given by

$$m = \frac{1}{2}Cn_c k = \frac{Cn_c}{2}[(n_c-1)p + n_c(C-1)q]. \tag{19}$$

We write down Q

$$\begin{aligned} Q &= \frac{C}{2m}\left[n_c(n_c-1)p - n_c^2\frac{\{(n_c-1)p+n_c(C-1)q\}^2}{2m}\right]\\ &= \frac{(n_c-1)p}{(n_c-1)p+n_c(C-1)q} - \frac{1}{C}. &(20)\end{aligned}$$

When $n_c \gg 1$ (which is always the case), we have that

$$Q = \frac{p}{p+q(C-1)} - \frac{1}{C}, \tag{21}$$

When we write q as some fraction f of p, that is $q = fp$, with $0 \leq f \leq 1$, we find

$$Q(C,f) = \frac{1}{1+(C-1)f} - \frac{1}{C}, \tag{22}$$

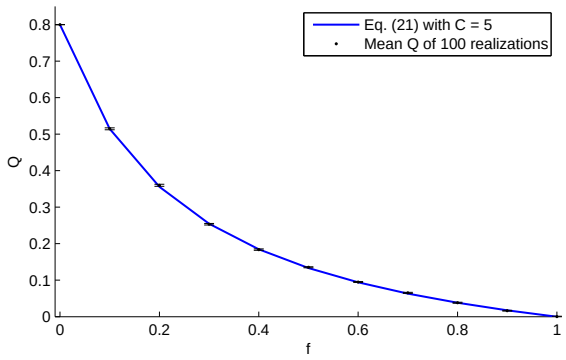


FIG. 2: Equation (22) and Q_{design}. This figure displays Q as a function of f (the relative probability of a link between communities), with $C = 5$ for the simple network defined in Figure 1. The blue line is given by Eq. (22) and the black dots with error-bars are mean values of Q_{design} in realizations of the simple network with $p = 1/10$ and $n = 500$; each data-point is the mean of 100 realizations. The error bars are calculated as the standard deviation divided by square root of the number of runs.

which is independent of p. Thus, for this simple network, the only two relevant parameters are the number of communities and the density of the inter-community links relative to the intra-community strength. We can also see that our result from Eq. (17) is valid even in the case $p < 1$, as long as the communities are connected and $q = 0$.

If we design an adjacency matrix according to Figure 1, we can calculate the value $Q_{\text{design}} = \text{Tr}(\mathbf{S}_{\text{d}}^T \mathbf{B} \mathbf{S}_{\text{d}})/(2m)$, where $\mathbf{S}_{\text{d}}$ is a community-matrix that reflects the designed communities. Values of Q_{design} should correspond to Eq. (22). We see in Figure 2 that this expectation is indeed fulfilled. The blue curve is Q as a function of f with $C = 5$. The black dots with error-bars are mean values of Q_{design} in realizations of the simple network with $p = 1/10$ and $n = 500$; each data-point is the mean of 100 realizations and the error bars are calculated as the standard deviation divided by square root of the number of runs. The correspondence between prediction and experiment is quite compelling.

We should note, however, that the value of Q_{design} may be lower than the actual modularity found for the network by a good algorithm: We can imagine that fluctuations of the inter-community links could result in configurations that would yield higher values of Q—especially for high values of f. We can quantify this quite precisely. Reichardt and Bornholdt [16] have shown that demonstrated that random networks can display significantly larger values of Q due to fluctuations; when $f = 1$, our simple network is precisely a random network (see also related work by Guimerà *et al.* [21]). In the case of the network we are experimenting on, ($n = 500$, $p = 1/10$), they predict $Q \approx 0.13$.

Thus, we expect that the curve for $Q(f,C)$ with fixed C will be deviate from the Q_{design} displayed in Figure 2; especially for values of f that are close to unity. The line will decrease monotonically from $Q(0,C) = 1 - 1/C$ towards $Q(1,C) = 0.11$ with the difference becoming maximal as $f \to 1$.

VI. NUMERICAL EXPERIMENTS

We know that the running time of mean field method scales like that of the spectral solution. In order to compare the precision of the mean field solutions to the solutions stemming from spectral optimization, we have created a number of test networks with adjacency matrices designed according to Figure 1. We have created 100 test networks using parameters $n_c = 100$, $C = 5$, $p = 0.1$ and $f \in [0,1]$. Varying f over this interval allows us to interpolate between a model with C disjunct communities and a random network with no community structure.

We applied the following three algorithms to our test networks

1. Spectral optimization,
2. Spectral optimization and the KLN-algorithm, and
3. Mean field optimization.

Spectral optimization and the KLN-algorithm were implemented as prescribed in [11]. The nC non-linear mean field annealing equations were solved approximately using a $D = 300$-step *annealing schedule* linear in $\beta = 1/T$ starting at β_c and ending in $3\beta_c$ at which temperature the majority of the mean field variables are saturated. The mean field critical temperature $T_c = b_{\max}/C$ is determined for each connectivity matrix. The synchronous update scheme defined as parallel update of all means at each of the D temperatures

$$\begin{aligned} \mu_{ik}^{(d+1)} &= \frac{\exp(\phi_{ik}^{(d)}/T)}{\sum_{k'=1}^{C} \exp(\phi_{ik'}^{(d)}/T)} \\ \phi_{ik}^{(d)} &= \sum_j B_{ij} \mu_{jk}^{(d)} \end{aligned} \tag{23}$$

can grow unstable at low temperatures. A slightly more effective and stable update scheme is obtained by selecting random fractions $\rho < 1$ of the means for update in $1/\rho$ steps at each temperature. We use $\rho = 0.2$ in the experiments reported below. A final $T = 0$ iteration, equivalent to making a decision on the node community assignment, completes the procedure. We *do not* assume that actual the number of communities $C < C_{\max}$ is known in advance. In these experiments we use $C_{max} = 8$. This number is determined after convergence by counting the number of non-empty communities

The results of the numerical runs are displayed in Figure 3. This figure shows the point-wise differences between the value of $Q_{\text{algorithm}}$ found by the algorithm in question and Q_{design} plotted as a function of the inter-community noise f. The line of $Q_{\text{algorithm}} - Q_{\text{design}} = 0$ thus corresponds to the curve plotted in Figure 2. We see from Figure 3 that the mean field approach uniformly out-performs both spectral optimization and spectral optimization with KLN post-processing. We also ran a Gibbs sampler [16] for with a computational complexity equivalent to the mean field approach. This lead to communities with Q slightly lower than the mean field results, but still better than spectral optimization with KLN post-processing.

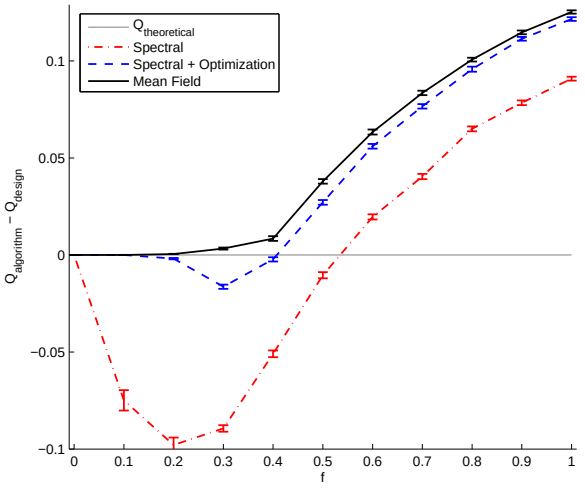


FIG. 3: Comparing spectral methods with the mean field solution. The networks were created according to the simple model, using parameters $n_c = 100$, $C = 5$, $p = 0.1$ and $f \in [0,1]$. All data points display the point-wise differences between the value of $Q_{\text{algorithm}}$ found by the algorithm in question and Q_{design}. The error-bars are calculated as in Figure 2. The dash-dotted red line shows the results for the spectral method. The dashed blue line shows the results for the spectral optimization followed by KLN post-processing. The solid black curve shows the results for the mean field optimization. The grey, horizontal line corresponds to the theoretical prediction (Eq. (22)) for the designed communities.

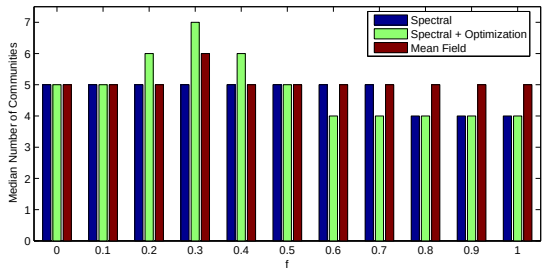


FIG. 4: The median number of communities found by the various algorithms. The panel shows the median number of communities as a function of the relative fraction of inter-community links f. All optimization schemes consistently pick four or five communities for the highest values of f. This finding is consistent with theoretical and experimental results by Reichardt and Bornholdt [16]

We note that the obtained $Q_{\text{algorithm}}$ for a random network ($f = 1$) is consistent with the prediction made by Reichardt and Bornholdt [16]. We also see that the optimization algorithms can exploit random connections to find higher values of $Q_{\text{algorithm}}$ than expected for the designed communities Q_{design}. In the case of the mean field algorithm this effect is visible for values of f as low as 0.2.

Figure 4 shows the median number of communities found by the various algorithms as a function of f. It is evident from Figs. 3 and 4 that—for this particular set of parameters—the problem of detecting the designed community structure is especially difficult around $f = 0.3$. Spectral clustering with and without the KLN algorithm find values $Q_{\text{algorithm}}$ that are significantly lower than Q_{desigm}. The mean field algorithm manages to find a value of $Q_{\text{algorithm}}$ that is higher than the designed Q but does so by creating extra communities. As $f \to 1$ it becomes more and more difficult to recover the designed number of communities.

VII. CONCLUSIONS

We have introduced a deterministic mean field annealing approach to optimization of modularity Q. We have evaluated the performance of the new algorithm within a family of networks with variable levels of inter-community links, f. Even with a rather costly post-processing approach, the spectral clustering approach suggested by Newman is consistently out-performed by the mean field approach for higher noise levels. Spectral clustering without the KLN post-processing finds much lower values of Q for all $f > 0$.

Speed is not the only benefit of the mean field approach. Another advantage is that the implementation of mean field annealing is rather simple and similar to Gibbs sampling. This method also avoids the inherent problems of repeated bisection. The deterministic annealing scheme is directed towards locating optimal configurations without wasting time at careful thermal equilibration at higher temperatures. As we have noted above, the modularity measure Q may need modification in specific non-generic networks. In that case, we note that the mean field method is quite general and can be generalized to many other measures.

[1] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004, cond-mat/0308217.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74:47, 2002.

[3] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51:1079, 2002.

[4] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[5] M.E.J. Newman. Detecting community structure in networks. *The European Physical Journal B*, 38:321, 2004.

[6] L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics*, page P09008, 2005, cond-mat/0505245.

[7] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104, 2006.

[8] A.-L. Barabási and Z. N. Oltvai. Network biology: Understand-

ing the cell's functional organization. *Nature Reviews Genetics*, 5:101, 2004.

[9] F. K. R. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[10] O. Goldscmidt and D. S. Hochbaum. Polynomial algorithm for the k-cut problem. In *Proceedings of the 29th Annual IEEE Symposium on the Foundations of Computer Science*, page 444. Institute of Electrical and Electronics Engineers, 1988.

[11] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, USA*, 103:8577, 2006.

[12] S. Fortunato and M. Barthelemy. Resolution limit in community detection. 2006, physics/0607100.

[13] M. Rosvall and C. T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. 2006, physics/0612035.

[14] Mathworld. http://mathworld.wolfram.com/.

[15] M.E.J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, 2004, cond-mat/0309508.

[16] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74:016110, 2006.

[17] B. M. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, 49:291, 1970.

[18] S. Kirkpatrick, C.D. Gelatt Jr., , and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[19] C. Peterson and J.R. Anderson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

[20] C. Peterson and B. Söderberg. A new method for mapping optimization problems onto neural networks. *Int J Neural Syst*, 1:3–22, 1989.

[21] R. Guimerá, M. Sales-Pardo, and L. A. N. Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70:025101, 2004, cond-mat/0403660.

Part III

Perspectives and Bibliography

CHAPTER 7

Perspectives

"Physicists, it turns out, are almost perfectly suited to invading other people's disciplines, being not only extremely clever but also generally much less fussy than most about the problems they choose to study. Physicists tend to see themselves as lords of the academic jungle, loftily regarding their own methods as above the ken of anybody else and jealously guarding their own terrain. But their alter egos are closer to scavengers, happy to borrow ideas and techniques from anyone if they seem like they might be useful, and delighted to stomp all over someone else's problem. As irritating as this attitude can be to everybody else, the arrival of physicists into a previously non-physics area of research often presages a period of great discovery and excitement."

— Duncan J. Watts, *Six Degrees* [93]

WELL, the scavengers are certainly here—that much is evident from the ubiquitousness of physicists and ex-physicists in the bibliography of this dissertation. — And Watts really gets it right: The science of networks has indeed experienced a period of great discovery and excitement. Luckily there is still a great number of juicy subjects left to scavenge.

This concluding chapter summarizes the work presented in the previous chapters and points out the subjects that I consider particularly interesting and ripe for future research.

7.1 Scientific Citations

Much of the work in this dissertation regards the network of scientific citations. Two papers describe modelling of the degree distributions of live- and dead papers in the SPIRES data base [53, 54]. The distinction between live and dead nodes is based on the fact that many nodes in the SPIRES network are 'dead' in the sense that they have not been cited in many years and that they will (most likely) never be cited again. The proposed model is an augmentation of the simple growth model proposed by Barabási and Albert [10] (cf. section 1.4), where, at each update, in addition to the standard probability of gaining a new link, each node of the network also has a finite probability (inversely proportional to the number of links) of becoming permanently inactive (i.e. dead). The model is analytically soluble and the solution provides a surprisingly good fit to the empirical distributions of live and dead papers in the data base, with only three parameters (mean degree of live distribution, mean degree of dead distribution, and the fraction of dead papers).

The live/dead model is particularly interesting because many other networks possess nodes that can become permanently inactive due to age. Further, it is demonstrated that the death-mechanism alone can result in power-law degree distributions.

Citation data is used to evaluate individual scientists. This is the starting point for the two papers [56, 57] that investigate the author-network in the SPIRES data base. In particular, authors (the nodes in figure 3.2) are characterized by their publication record $\{k_i\}$, which is a list of in-degrees from the paper-network of their publications.

Lehmann *et al.* use Bayesian statistics to analyze several measures of scientific quality. More precisely, each measure's ability to discriminate between authors is determined. By use of scaling arguments, it is shown that the best of these

measures (the mean and median) require approximately 50 papers in order to draw conclusions regarding long-term scientific performance with usefully small uncertainties. The probabilistic methods also permit comparison of scientists working in different areas.

That approximately 50 publications are required to make author assignments to decile groups with 90% certainty is large enough to limit the utility of such analyses in the academic appointment process. Therefore, improvement of this number is an interesting subject for future research. Are there more efficient measures of an author's full citation record than those considered in [56, 57]?

- First of all, preliminary investigations have revealed that straightforward iterative schemes can be constructed to create more efficient measures. The iterative schemes converge rapidly to an 'optimal' binning of authors into deciles. The result is optimal in the sense that it maximizes the KL divergences between the conditional distributions $P(i|\alpha)$ (see section 5.2). The results for these optimal binnings, however, appear to be only marginally better than those obtained for the mean or median measures. This work is in progress.

- Secondly, it is important to recognize that it takes time for a paper to accumulate its full complement of citations. Therefore, co-author linkages may provide a powerful supplement or alternative to citation data, since the list of co-authors remains unaltered from the moment a paper is submitted for publication. We can empirically construct co-author distributions conditioned on any measure of quality, and analyze these completely analogously, with the analysis of citation data presented in chapter 5. Preliminary studies of the probability that authors in bins α and β will co-author a publication reveal a striking concentration along the diagonal $\alpha = \beta$. This promising line of work is also in progress. It appears that we may be able to substantially reduce the number of papers needed in order to assign authors to bins with a 90% accuracy level.

- Thirdly, we know that the author network possesses a high degree of stratification. Until now, we have made the assumption that all authors in the data base pick their topics and give their references with as much care and

insight as the top 10%. Assume for a moment that the ability of authors to identify the 'right' papers varies as much as—and is directly related to—the ability to write highly cited papers.

If this is the case, the roughly 50% of references that the lower half of authors contribute, could be essentially noise with no useful information about the quality of papers. We may obtain a much sharper measure by considering only the references given by the top 50% of authors. Excluding these citations disqualifies approximately half of the data, but if the data we lose is mostly noise rather than signal, it can only improve things. Implementing the assumption of 'variation in the ability to give references' into the analysis of citations is also a work in progress.

A related project stems from the fact that the citation network is an information network. Let us assume that the number of citations pointing towards a given author is proportional to the amount of information, he has supplied to the network. Let us make the related assumption that each reference is similarly proportional to the amount of information received from the network. Since we only consider links that are internal to SPIRES, there is a conservation of information.

With these assumptions, we can explicitly study the flow of information between authors in SPIRES. Preliminary investigations of the flow of information between the decile groups of authors based on the mean number of citations, have been performed. The only group that has an out-flux of information is the group containing the ten percent highest cited authors. Group nine breaks approximately even and the remaining 80 percent of authors receive information. This analysis emphasizes the conclusion from [58] that only a select few authors drive the progress in theoretical high energy physics. It also points towards interesting results for investigations of the directed author-network with links weighted by information flux, cf. figure 3.3.

In summary, a statistical approach is highly useful when the network contains more structure than what can be contained in an adjacency matrix. Since we would like to include information about the nodes into the analysis of many other networks (social networks, the internet, etc.), this approach may find wider use in the future.

7.2 Communities and Beyond

The paper *Deterministic Community Detection* [52] discusses the use of mean field methods to maximize the modularity Q. In order to test the mean field algorithm in a controlled environment, Lehmann and Hansen constructed a simple class of random networks with adjustable community structure and an analytic expression for the designed modularity Q. Using these networks, Lehmann and Hansen show that the mean field methods displays a better performance than previously known deterministic (e.g. spectral) methods.

In the context of optimization of Q, one interesting topic for future research is an investigation of relationship between the deterministic methods, such as the mean field algorithm, and probabilistic methods, such as Gibbs sampling. For the networks we have investigated, the two methods performance comparably in terms of values of Q. However, if the community structure is fuzzy, the two methods find dissimilar communities for the same network. Understanding why this is the case, will lead to a deeper understanding of both the mean field and probabilistic methods.

In a more general setting, we would like to develop better measures for community structure. Imagine a network with power-law degree distributions *and* community structure. We can imagine different scenarios for the hubs of such a network

(i) *Hubs are 'shielded' from the remaining network by their communities.* One example of a network, where hubs are possibly shielded from the remaining network, is the network of academic co-authorship. Professors are the most connected nodes (hubs), but they link mostly to post. docs. and grad-students in their own department. The two latter groups are more mobile and link more often to other research groups. At the same time, however, they 'insulate' professors from the rest of the network.

Of course the role of nodes change over time in this network. At one point in time, the professors themselves have been both grad-students and post. docs.

(ii) *Hubs lie on the boundary between two communities.* This situation is proba-

bly the case in the network of musical performers linked by co-play in user playlists. Most people stick to one or two genres of music (communities) that they like, but most people also like to listen to hit songs. The performers behind these songs span several genres and become hubs that connect different communities: People who listen mostly to pop may still play an occasional song by the hub *Metallica*, and people who are dedicated to hard rock, probably like the song 'Lose Yourself', even if it is written and performed by the rapper *Eminem*.

So what can we learn from these observations? First of all, if we try to visualize the structure of these communities, case (i) should remind us of figure 1.9 a), that is: In the case of shielded hubs, each community, must possess a high level of topological hierarchy. Similarly, we may expect that communities, where hubs lie on the boundary, are somewhat anti-hierarchically structured. In case (ii), it also becomes extremely important that our community detection algorithm is able to handle overlap between modules in a natural way.

Our ambition with respect to measures of community structure is to develop a measure that is able to handle communities in power-law networks, e.g. by considering the level of topological hierarchy in each community. One could utilize the level of hierarchy in each community to estimate the possible presence of hubs on the boundary of that module; in this way, one could assign the hubs to each of their various communities.

Let me end this dissertation with a short poem by the Danish[1] writer, Henrik Nordbrandt:

> Søg ingen sandhed her. Disse digte er håndens værk
> som den bevægede sig nogle dage i november, eller skælvede
> påvirket af sin ejers humør, kaffe, cigaretter, vin
> skyerne over dalen, venners død og meddelelser om krige.
>
> — Henrik Nordbrandt, *Håndens skælven i november* (1986)

[1] I apologize to those readers who do not understand Danish.

Bibliography

[1] L. A. Adamic and B. A. Huberman. Power-law distribution of the world wide web. *Science*, 287:2115, 2000.

[2] L. A. Adamic, R. M. Lukose, and B. A. Huberman. *Handbook of Graphs and Networks*, chapter Local search in unstructured networks. Wiley–VHC, Berlin, 2003.

[3] L. A. Adamic, R. M Lukose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *Physical Review E*, 64:046135, 2001.

[4] R. Albert, I. Albert, and G. L. Nakarado. Structural vulnerability of the north american power grid. *Physical Review E*, 69:025103, 2004.

[5] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74:47, 2002.

[6] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world wide web. *Nature*, 401:130, 1999.

[7] R. Albert, H. Jeong., and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378, 2000.

[8] A.-L. Barabási. *Linked: The New Science of Networks*. Perseus, 2002.

[9] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[10] A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: The topology of the world wide web. *Physica A*, 281:69, 2000.

[11] A.-L. Barabási, R. Albert, H. Jeong, and G. Bianconi. Power-law distribution of the world wide web. *Science*, 287:2115a, 2000.

[12] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424:175, 2006.

[13] S. Bornholdt and H. G. Schuster, editors. *Handbook of Graphs and Networks*. Wiley–VHC, Berlin, 2003.

[14] J. S. A. Bridgewater, O. P. Boykin, and V. P. Roychowdhury. Statistical mechanical load balancer for the web. *Physical Review E*, 71:046133, 2005.

[15] S. Brin and L. Page. The anatomy of a large scale hypertextual web search engine. *Computer Networks*, 30:107, 1998.

[16] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, and A. Tomkins. Graph structure in the web. *Computer Networks*, 33:309, 2000.

[17] A. Clauset, M.E.J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[18] R. Cohen and S. Havlin. Scale-free networks are ultrasmall. *Physical Review Letters*, 90:058701, 2003.

[19] L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics*, page P09008, 2005.

[20] D. de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27:292, 1976.

[21] I. Derényi, G. Palla, and T. Vicsek. Clique percolation in random networks. *Physical Review Letters*, 94:160202, 2005.

[22] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51:1079, 2002.

[23] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72:027104, 2005.

[24] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 64:035103, 2002.

[25] P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290, 1959.

[26] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23:298, 1973.

[27] The National Laboratory for Applied Network Research. Data from the Oregon Views project. http://moat.nlanr.net/AS/.

[28] S. Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104:36, 2007.

[29] B. Foxman, M. E. J. Newman, B. Percha, K. K. Holmes, and S. O. Aral. Measures of sexual partnerships: Lengths, gaps, overlaps and sexually transmitted infection. *Sexually Transmitted Diseases*, 33:209, 2006.

[30] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35, 1977.

[31] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences USA*, 99:7821, 2002.

[32] O. Goldschmidt and D. S. Hochbaum. Polynomial algorithm for the k-cut problem. In *Proceedings of the 29th Annual IEEE Symposium on the Foundations of Computer Science*, page 444, New York, 1988. Institute of Electrical and Electronics Engineers.

[33] S. Guattery and G. L. Miller. On the quality of spectral separators. *Siam Journal of Matrix Analysis and Applications*, 19:701, 1998.

[34] R. Guimerá, M. Sales-Pardo, and L. A. N. Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70:025101, 2004.

[35] S. Gupta, R. M. Anderson, and R. M. May. Networks of sexual contacts: Implications for the pattern of spread of HIV. *AIDS*, 3:807, 1989.

[36] K. B. Hajra and P. Sen. Modelling aging characteristics in citation networks. *Physica A*, 368:575, 2006.

[37] J. E. Hirsch. An index to quantify an individual's scientific output. *Proceedings of the National Academy of the Sciences*, 102:16569, 2005.

[38] A. Jaffe and M. Trajtenberg. *A Window on the Knowledge Economy*. MIT Press, Cambridge, MA, 2002.

[39] E. T. Jaynes. *Probability Theory: The Logic of Science: Principles and Elementary Applications. Vol. 1*. Cambridge University Press, 2003.

[40] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31:7, 1988.

[41] B. J. Kim, C. N. Yoon, S. K. Han, and H. Jeong. Path finding strategies in scale-free networks. *Physical Review E*, 65:027103, 2002.

[42] O. Kinouchi, A. S. Martinez, G. F. Lima, G. M Lourenco, and S. Risau-Gusman. Deterministic walks in random networks: An application to thesaurus graphs. *Physica A*, 315:665, 2002.

[43] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:604, 1999.

[44] J. M. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The Web as a graph: Measurements, models and methods. In *Lecture Notes in Computer Science*, volume 1627, page 1. Springer, 1999.

[45] D. E. Knuth. *The Stanford GraphBase: A Platform for Combinatorial Computing*. Addison-Wesley, Reading, MA, 1993.

[46] R. N. Kostoff. *Science and Technology Metrics*. Storming Media, 2005.

[47] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63:066123, 2001.

[48] P. L. Krapivsky and S. Redner. Network growth by copying. *Physical Review E*, 71:036118, 2005.

[49] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Physical Review Letters*, 85(21):4629, 2000.

[50] H. W. Lauw, E.-P. Lim, H. Pang, and T.-T. Tan. Social network discovery by mining spatio-temporal events. *Computational & Mathematical Organization Theory*, 11:97, 2005.

[51] S. Lehmann. Spires on the building of science. Master's thesis, The Niels Bohr Institute, 2003.

[52] S. Lehmann and L. K. Hansen. Deterministic modularity optimization. *physics/0701348*, 2007. Submitted to Physical Review E.

[53] S. Lehmann and A. D. Jackson. Live and dead nodes. *Computational & Mathematical Organization Theory*, 11:161, 2005.

[54] S. Lehmann, A. D. Jackson, and B. Lautrup. Life, death and preferential attachment. *Europhysics Letters*, 69:298, 2005.

[55] S. Lehmann, A. D. Jackson, and B. E. Lautrup. Measures and mismeasures of scientific quality. *physics/0512238*, 2005.

[56] S. Lehmann, A. D. Jackson, and B. E. Lautrup. Measures for measures. *Nature*, 444:1003, 2006.

[57] S. Lehmann, A. D. Jackson, and B. E. Lautrup. A quantitative analysis of measures of quality in science. *physics/0701311*, 2007. Submitted to Physical Review E.

[58] S. Lehmann, B. E. Lautrup, and A. D. Jackson. Citation networks in high energy physics. *Physical Review E*, 68:026113, 2003.

[59] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910, 2002.

[60] S. Maslov, K. Sneppen, and A. Zaliznyak. Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A*, 333:529, 2004.

[61] S. Milgram. The small world problem. *Psychology Today*, 2:60, 1967.

[62] R. Milo, S. S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824, 2002.

[63] A. E. Motter, A. P. de Moura, Y.-C. Lai, and P. Dasgupta. Topology of the conceptual network of language. *Physical Review E*, 65:065102, 2002.

[64] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89:208701, 2002.

[65] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[66] M. E. J. Newman. Detecting community structure in networks. *The European Physics Journal B*, 38:321, 2004.

[67] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323, 2005.

[68] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104, 2006.

[69] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, USA*, 103:8577, 2006.

[70] M. E. J. Newman. Mathematics of networks. In *The New Palgrave Encyclopedia of Economics*. Palgrave Macmillan, second edition, In Press.

[71] M. E. J. Newman, A.-L. Barabási, and D. J. Watts. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[72] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118, 2001.

[73] M.E.J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, 2004.

[74] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.

[75] H. O'Connell. Physicists thriving with paperless publishing. *High Energy Physics Libraries Webzine*, 2002.

[76] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814, 2005.

[77] G. Palla, I. Derényi, and T. Vicsek. The critical point of *k*-clique percolation in the Erdös-Rényi graph. *Journal of Statistical Mechanics*, 2006.

[78] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the internet. *Physical Review Letters*, 87:258701, 2001.

[79] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86:003200, 2001.

[80] A. Pothen, H. Simon, and K.-P. Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications*, 11:430, 1990.

[81] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of the Sciences, USA*, 101:2658, 2004.

[82] S. Redner. Citation statistics from more than a century of physical review. In *APS March Meeting*. American Physical Society, 2005.

[83] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74:016110, 2006.

[84] M. Rosvall and C. T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *physics/0612035*, 2006.

[85] S. Sanyal. Effect of citation patterns on network structure. *physics/0611139*, 2006.

[86] W. Shakespeare. *Measure for Measure*. Arden Shakespeare, 1967.

[87] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *escherichia coli*. *Nature Genetics*, 31:64, 2002.

[88] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888, 2000.

[89] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425, 1955.

[90] D. A. Spielman and S.-H. Teng. Spectral partitioning works: Planar graphs and finite element meshes. In *Proceedings of the 37th Annual IEEE Conference on Foundations of Computer Science*, 1996.

[91] A. Trusina, S. Maslov, P. Minnhagen, and K. Sneppen. Hierarchy measures in complex networks. *Physical Review Letters*, 92:178702, 2004.

[92] A. Vlado. Pajek: Program for large network analysis. http://vlado.fmf.uni-lj.si/pub/networks/pajek/. Version 1.17.

[93] D. J. Watts. *Small Worlds*. Princeton University Press, Princeton, 1999.

[94] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440, 1998.

[95] Y.-C. Wei and C.-K. Cheng. Towards efficient. hierarchical designs by ratio cut partitioning. In *Proceedings of the IEEE International Conference on Computer Aided Design*, page 298, New York, 1989. Institute of Electrical and Electronics Engineers.

[96] E. W. Weisstein. Complete graph. From MathWorld—A Wolfram Web Resource, 2006. http://mathworld.wolfram.com/CompleteGraph.html.

[97] G. U. Yule. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. *Philosophical Proceedings of the Royal Society of London. Series B*, 213:21, 1925.